Contrast and Generation Make BART a Good Dialogue Emotion Recognizer

Shimin Li[1,3], Hang Yan[1,3], Xipeng Qiu[1,2,3*]

[1] School of Computer Science, Fudan University
[2] Peng Cheng Laboratory, Shenzhen, Guangdong, China
[3] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
{smli20, hyan19, xpqiu}@fudan.edu.cn

https://github.com/whatissimondoing/CoG-BART.

—— *AAAI 2022*

2022.11.15 • ChongQing

Leibniz-Institut
für Sozialwissenschaften

Reported by Yuyang Lai

1.Introduction

2.Method

3.Experiments

gesis
Leibniz-Institut
für Sozialwissenschaften

Forschungszentrum · Research Center
L3S

Introduction

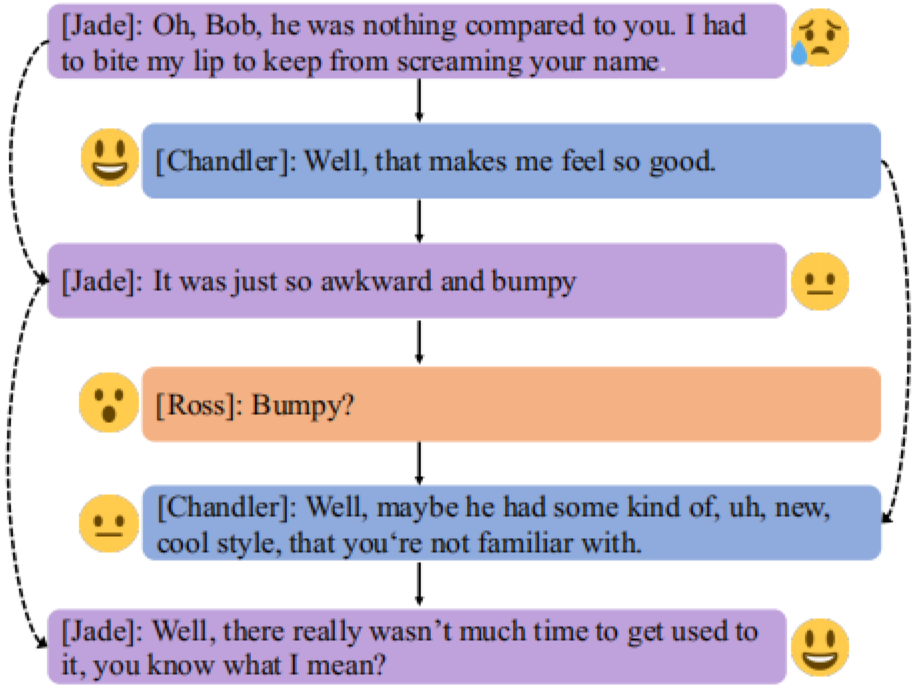


Figure 1: The conversation flow chart in multi-person dialogue emotion recognition.

- long-range contextual emotional relationships with speaker dependency.
- supervised contrastive learning
- auxiliary response generation task

Method

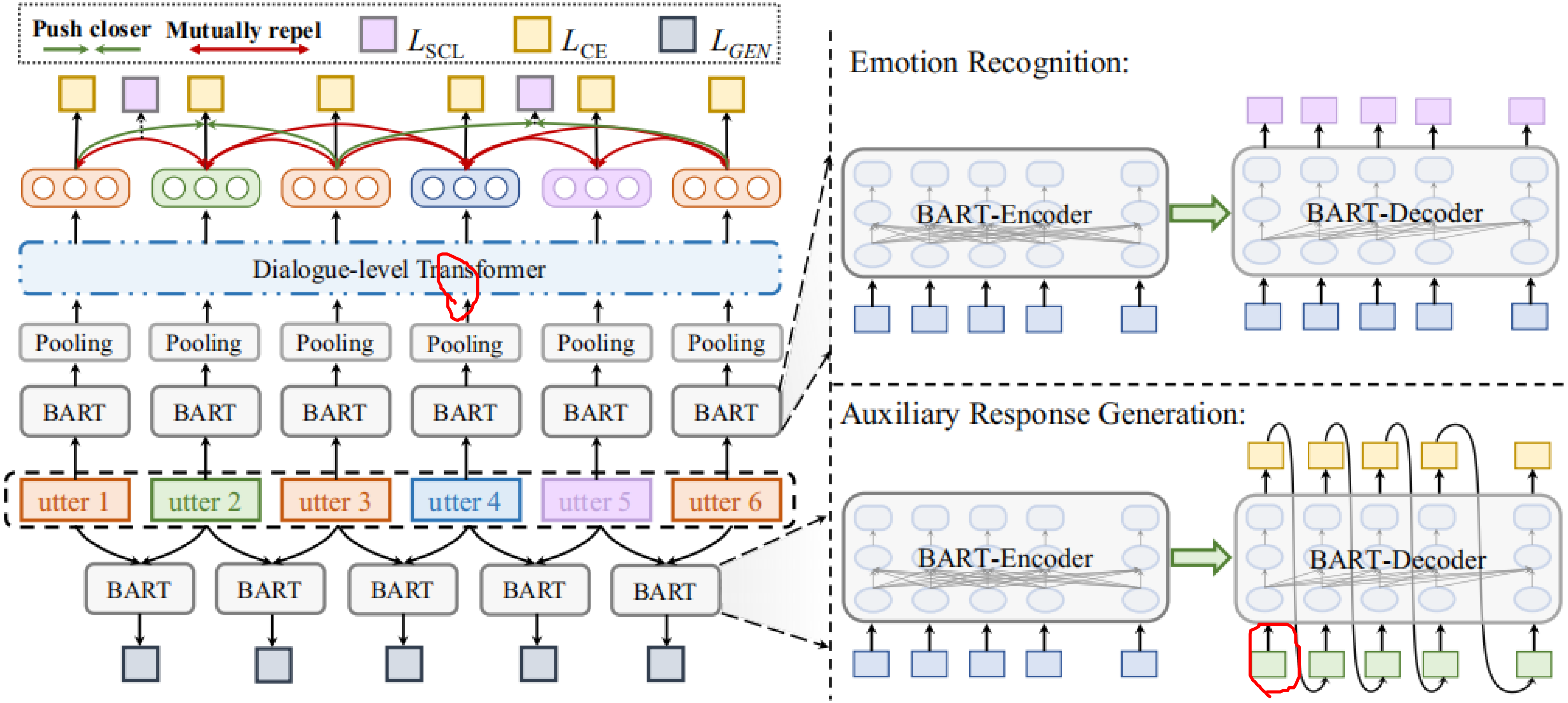
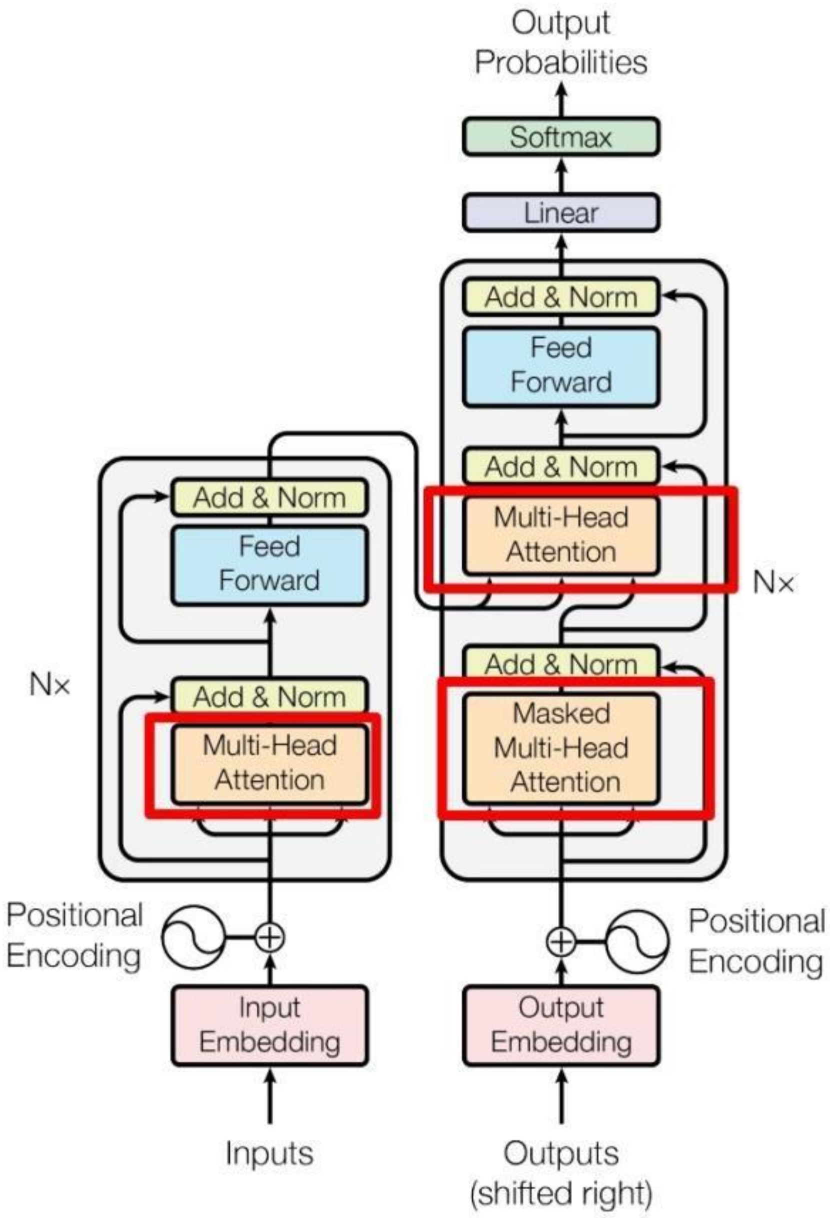
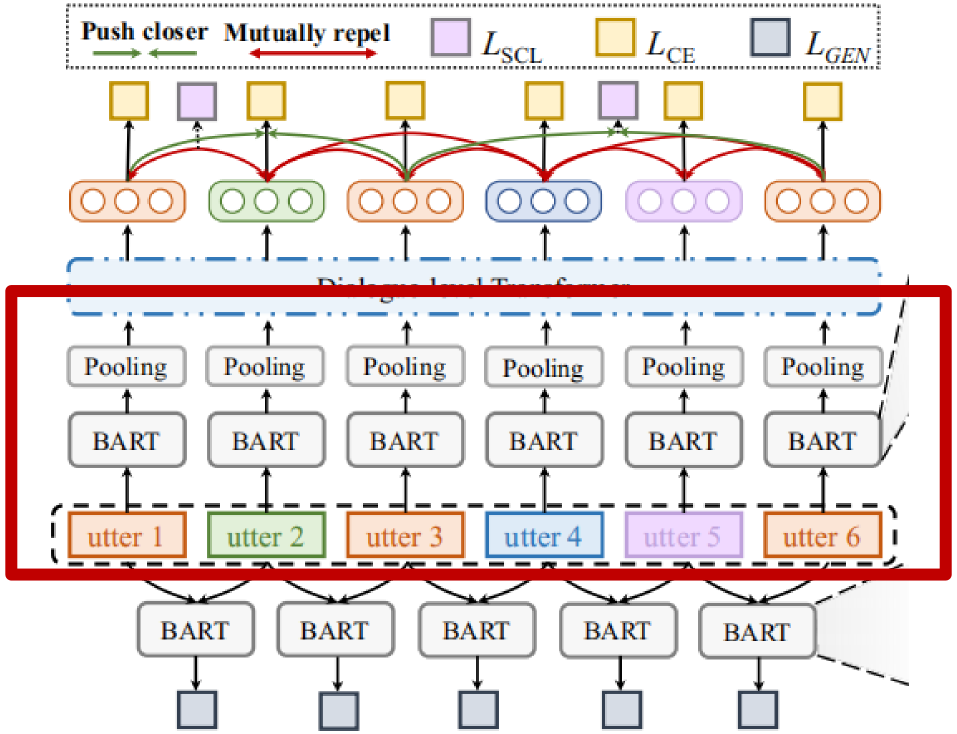


Figure 2: The overall framework of CoG-BART.



Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Method



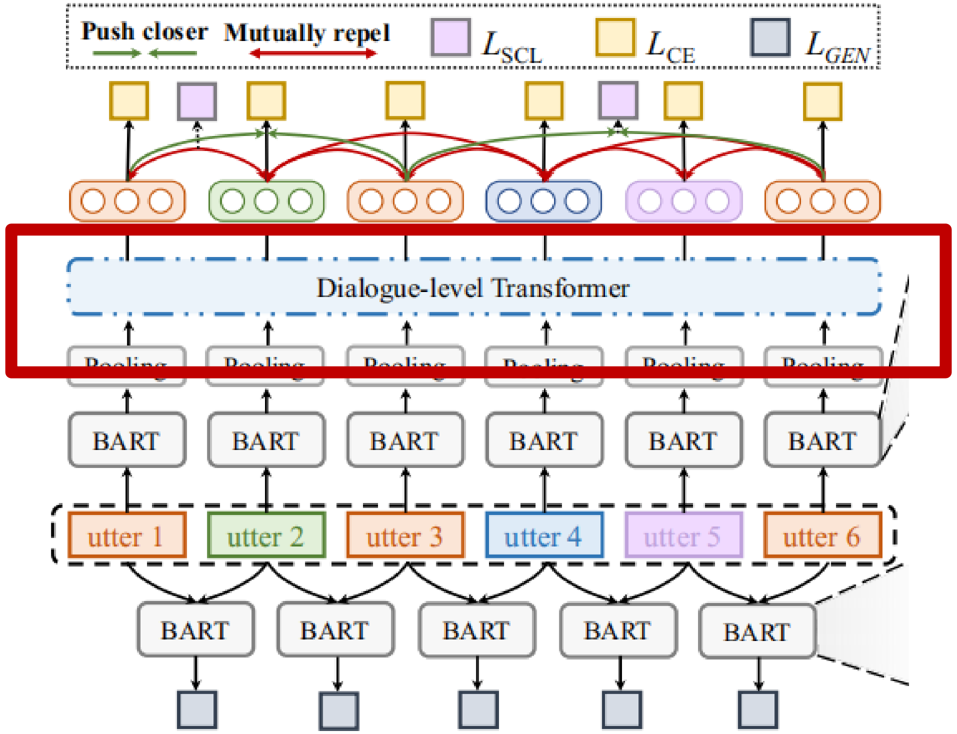
$$\tilde{u}_t = \left[\langle s\rangle, w_{t,1}, \cdots, w_{t,i}, \cdots, w_{t,|n_t|}, \langle/s\rangle\right], \quad (1)$$

$$H_t = \text{EmbeddingLayer}(\tilde{u}_t), \quad (2)$$

$$\widehat{H}_t = \text{BART-Model}(H_t), \quad (3)$$

$$\check{h}_t = \text{max-pooling}(\widehat{H}_t). \quad (4)$$

Method



$$\text{Atten}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{5}$$

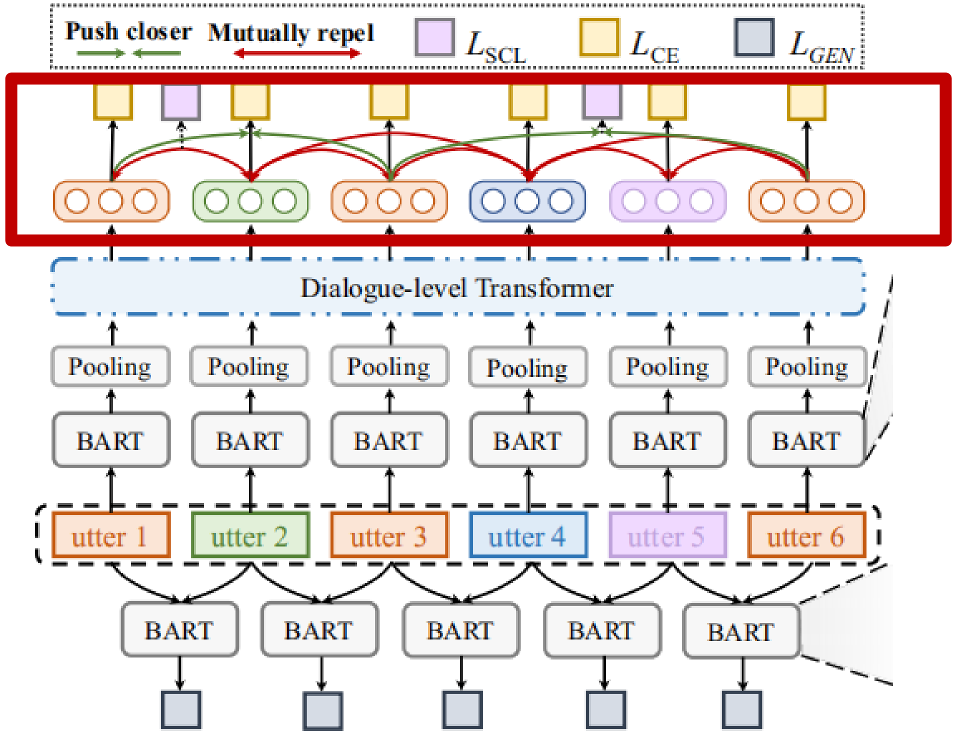
$$head_i = \text{Atten}(\check{h}_j W_i^Q, \check{h}_k W_i^K, \check{h}_k W_i^V), \tag{6}$$

$$\text{MultiHead}(Q, K, V) = [head_1; \cdots; head_n]W^O, \tag{7}$$

$$H_{win} = [\check{h}_t, \check{h}_{t+1}, \cdots, \check{h}_{t+bs-1}], \tag{8}$$

$$H_{d\text{-}win} = \text{Dialogue-Transformer}(H_{win}), \tag{9}$$

Method

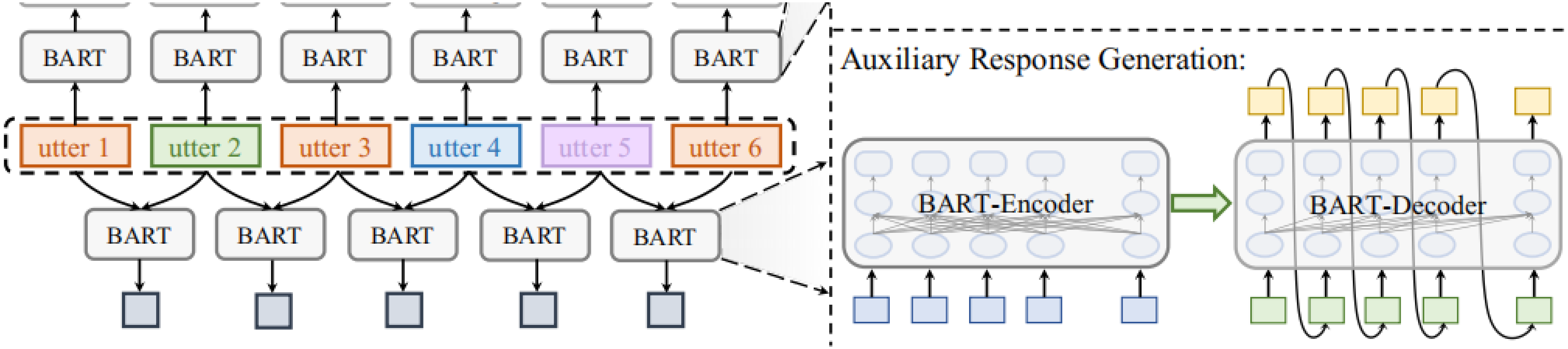


$$X = [H_{d\text{-}win}, \overline{H}_{d\text{-}win}], \tag{10}$$

$$\mathcal{L}_{\mathrm{SCL}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \mathrm{SIM}(p, i), \tag{11}$$

$$\mathrm{SIM}(p, i) = \log \frac{\exp((X_i \cdot X_p)/\tau)}{\sum_{a \in A(i)} \exp(X_i \cdot X_a/\tau)}, \tag{12}$$

Method



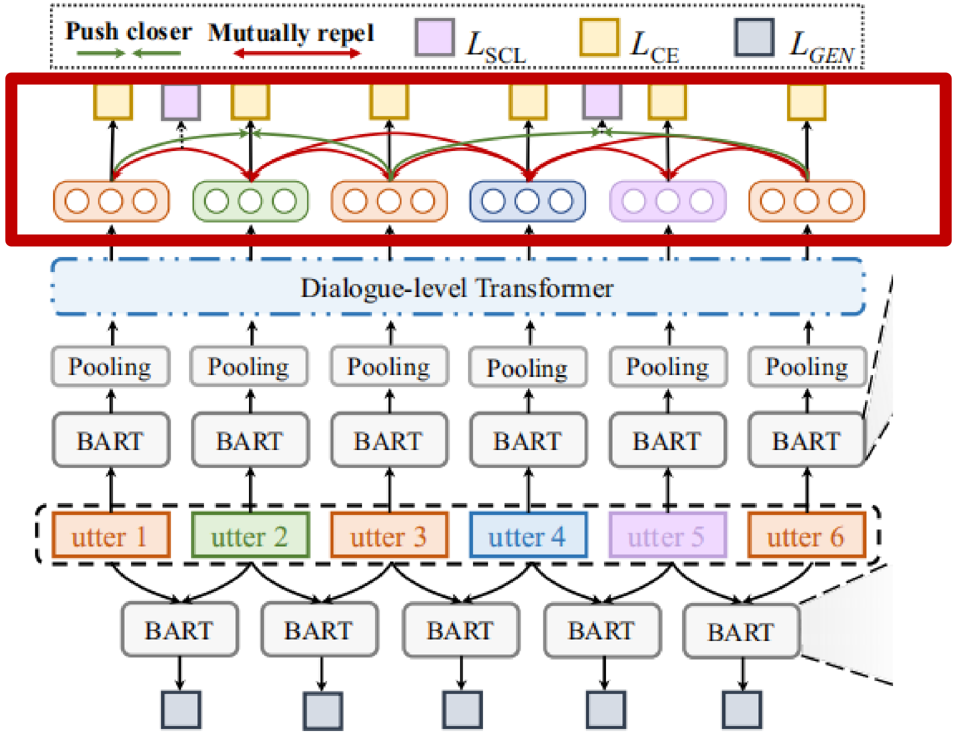
$$\acute{H}_t = \text{BART-Encoder}(H_t), \quad (13)$$

$$\grave{h}^d_j = \text{BART-Decoder}(\acute{H}_t; \grave{h}^d_{<j}), \quad (14)$$

$$u_{t+1,j} = \text{Softmax}(\grave{h}^d_j), \quad (15)$$

$$\mathcal{L}_{\text{Gen}} = -\sum_{t=1}^{N} \log p(u_{t+1} | u_t, \boldsymbol{\theta}), \quad (16)$$

Method



$$P_i = \text{Softmax}(W_s H_{d\text{-}win,i} + b_s), \tag{17}$$

$$\hat{y}_i = \text{argmax}(P_i), \tag{18}$$

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \cdot \log \hat{y}_{i,c}, \tag{19}$$

$$\mathcal{L} = (1-\alpha-\beta)\mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{SCL}} + \beta\mathcal{L}_{\text{Gen}}, \tag{20}$$

Experiments

Dataset		DD	MELD	ENLP	IEMOCAP
#Dial	Train	11118	1038	713	120
	Dev	1000	114	99	120
	Test	1000	280	85	31
#Utter	Train	87170	9989	9934	5810
	Dev	8069	1109	1344	5810
	Test	7740	2610	1328	1623
#CLS		7	7	7	6

Table 1: Statistics of four benchmark datasets.

Dataset	MELD		EmoryNLP		IEMOCAP		DailyDialog	
Model	Weighted -Avg-F1	Micro-F1	Weighted -Avg-F1	Micro-F1	Weighted -Avg-F1	Micro-F1	Weighted -F1-neural	Micro -F1-neutral
BERT	62.28	63.49	34.87	41.11	60.98	-	53.41	54.85
RoBERTa	62.51	63.75	35.90	40.81	63.38	-	52.84	54.33
HiTrans	61.94	-	36.75	-	64.50	-	-	-
DialogXL	62.41	-	34.73	-	65.94	-	-	54.93
XLNet	61.65	-	34.13	-	61.33	-	-	53.62
BART-large	63.57	64.41	35.98	38.93	56.14	56.67	54.83	55.34
CoG-BART	**64.81** ($\pm$0.19)	**65.95** ($\pm$0.44)	**39.04** ($\pm$0.10)	**42.58** ($\pm$0.94)	**66.18** ($\pm$0.45)	**66.71** ($\pm$0.49)	**56.09** ($\pm$0.01)	**56.29** ($\pm$0.17)

Table 2: The overall results of CoG-BART with pre-train-based baseline models on four datasets.

Experiments

Dataset		DD	MELD	ENLP	IEMOCAP
#Dial	Train	11118	1038	713	120
	Dev	1000	114	99	120
	Test	1000	280	85	31
#Utter	Train	87170	9989	9934	5810
	Dev	8069	1109	1344	5810
	Test	7740	2610	1328	1623
#CLS		7	7	7	6

Table 1: Statistics of four benchmark datasets.

Figure 3: The t-SNE visualization results of the model output when α is 0 and 0.8, respectively.

Dataset	MELD	EmoryNLP	IEMOCAP	DailyDialog
Model	Weighted -Avg-F1	Weighted -Avg-F1	Weighted -Avg-F1	Micro -F1-neutral
KET	58.18	34.39	59.56	53.37
RGAT	60.91	34.42	65.22	54.31
RGAT+RoBERTa	62.80	37.89	66.36	59.02
DialogGCN	58.10	-	64.18	-
DialogCRN	58.39	-	66.20	-
COSMIC	64.28	37.10	63.05	56.16
DAG-ERC	63.65	39.02	**68.03**	**59.33**
CoG-BART	**64.81** ($\pm$0.19)	**39.04** ($\pm$0.10)	66.18 ($\pm$0.45)	56.29 ($\pm$0.17)

Table 3: Comparison with graph-based models.

Metric	Weighted Average F1					
Datasets	α=0.2	α=0.4	α=0.6	α=0.8	β=0.1	β=0.2
MELD	**64.57**	63.99	64.42	61.84	64.83	63.70
IEMOCAP	64.38	**66.18**	65.12	63.38	66.18	63.54
EmoryNLP	**39.04**	36.68	36.90	35.24	37.45	37.57

Table 4: The F1 scores for different values of α and β

Experiments

Utterance for Prediction	Generated Response	Predict w/o RG	Predict with RG	Golden label
Joey : Thursday's clearly not good for ya, pick a day!	Sarah : So that's two boxes of the Holiday Macaroons. On behalf of the Brown Birds of America, I salute you.	anger	joy	joy
Joey: Man, that was great! Huh? Can you believe how long we threw that ball around?	Rachel : Yeah, it is amazing it lasted that long.	surprise	joy	joy

Figure 4: Case studies show that response generation enables the model to correctly predict the emotion based on context.

Experiments

Dataset	MELD	IEMOCAP
Methods	Weight-Avg-F1	
CoG-BART	64.81	66.18
-Gen	64.26 (↓0.55)	64.74 (↓1.44)
-SCL loss	64.28 (↓0.53)	64.23 (↓1.95)
-Speaker	64.14 (↓0.67)	55.41 (↓10.77)
-Gen, SCL loss	63.57 (↓**1.24**)	62.90 (↓3.28)
-SCL loss, Speaker	63.72 (↓1.09)	54.83 (↓**11.35**)
-Gen, Speaker	64.02 (↓0.79)	54.95 (↓11.23)
-Dialog-Trans	64.40 (↓0.41)	64.19 (↓1.99)

Table 5: Ablation study to evaluate the impact of different components on the overall performance of the model on MELD and EmoryNLP

Thank you!