Contrast and Generation Make BART a Good Dialogue Emotion Recognizer

Shimin Li, Hang Yan, Xipeng Qiu

Fudan University

2022.01.08

Task Introduction

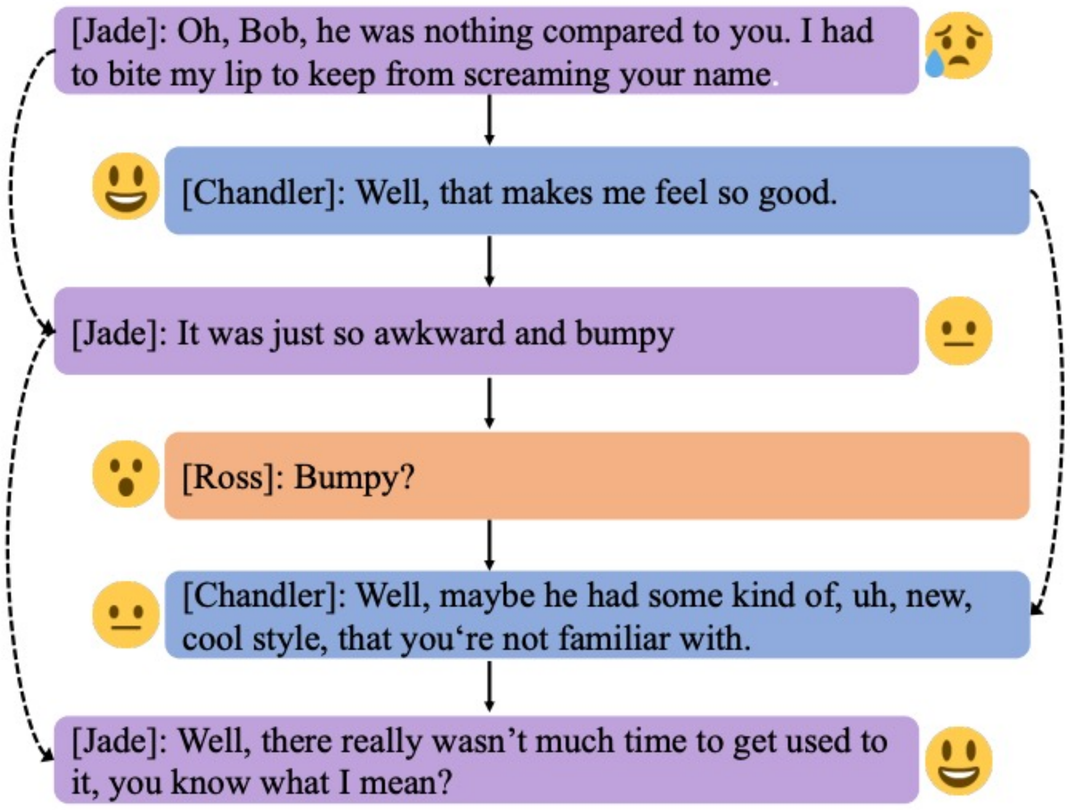
Emotion Recognition in Conversation:

- Identifying the emotion of several speakers' utterances in a conversation.

Social sentiment mining

empathetic dialogue system

Task Description

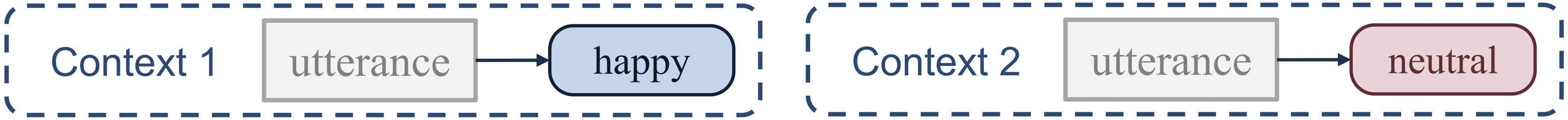


Characters:

- Chronological: the previous utterance directly influences the current speaker's emotion.
- the same speaker is influenced by other utterances and may expresses different emotions.

Main challenges

- The emotion of each utterance may be affected by contextual information.

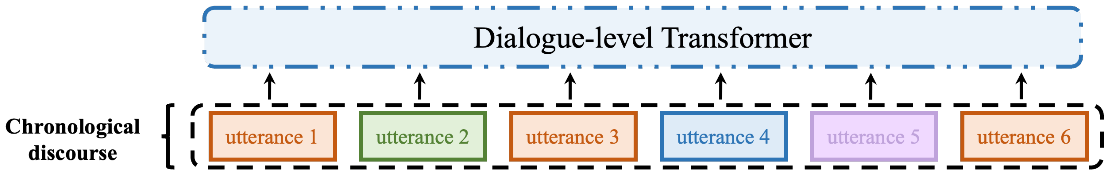


- Each speaker’s emotion is influenced by other speakers in the conversation (or emotional shifts).

- It is difficult to distinguish semantically similar sentiment categories, such as “frustrated” to “sad”, “happy” to “excited”, etc.

Method Overview

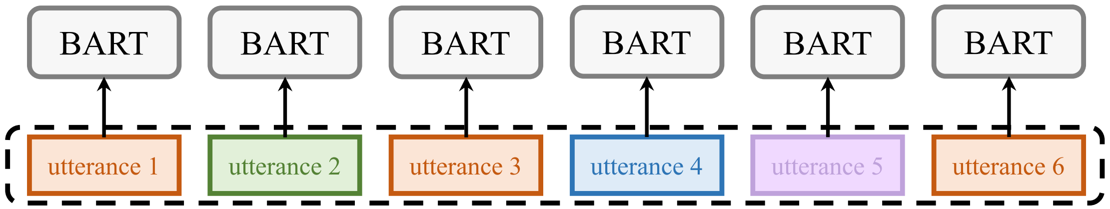
- To model context dependency:
  - Dialogue-level Transformer for long-range context dependency



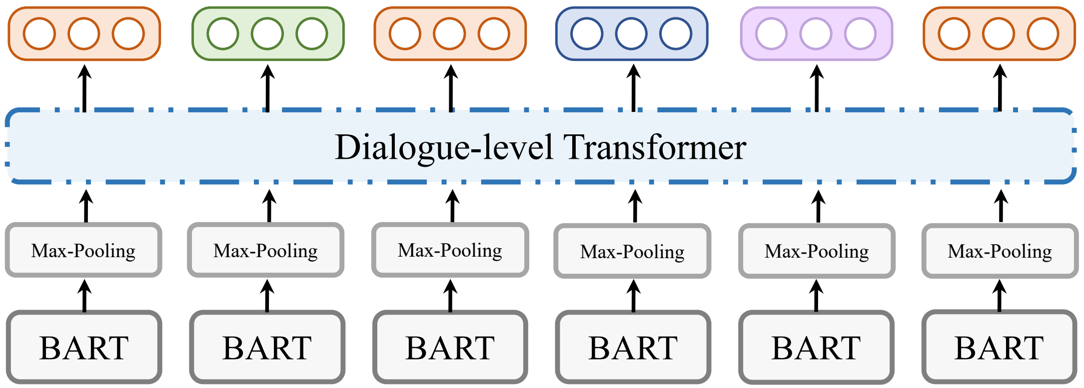
  - Response generation for short-turn context dependency
- To distinguish semantically similar emotions:
  - Supervised contrastive learning

Method

- Utterance Encoding:
  - ➢ splice the speaker's name or category before the utterance and encode with BART



- Dialogue Modeling:
  - ➢ modelling the contextual dependencies with dialogue-level Transformer for H_{d-win}



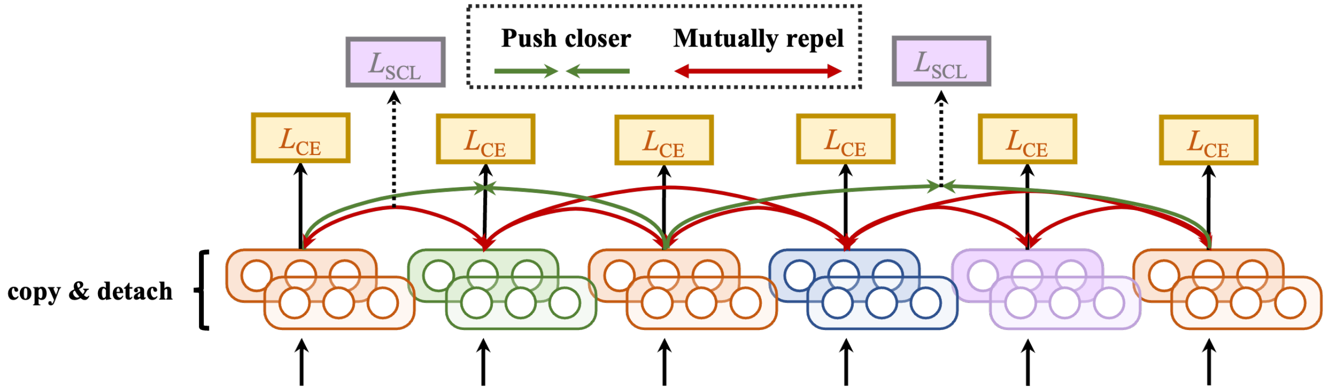
Method

- Supervised contrastive loss for ERC:
  - a copy of the hidden state of utterances is made to obtain $\overline{H}_{d-win}$, and its gradient is detached.

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

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

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

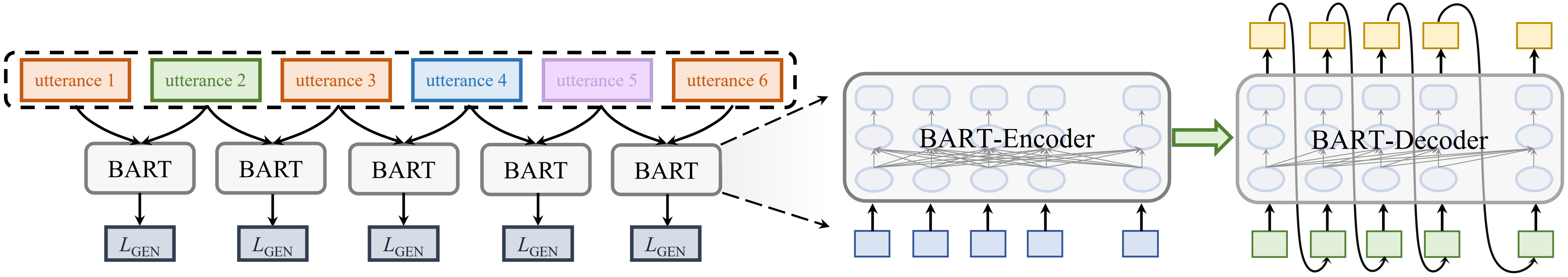


where $i \in I = \{1,2, \cdots, 2N\}$ indicate the index of the samples in a multi-view batch, $P(i) = I_{j=i} - \{i\}$ represents samples with the same category as i while excluding itself, $A(i) = I - \{i, N + i\}$ indicates samples in the multi-view batch except itself.

Method

- Auxiliary Response Generation
  - Generate response base on the current utterance.

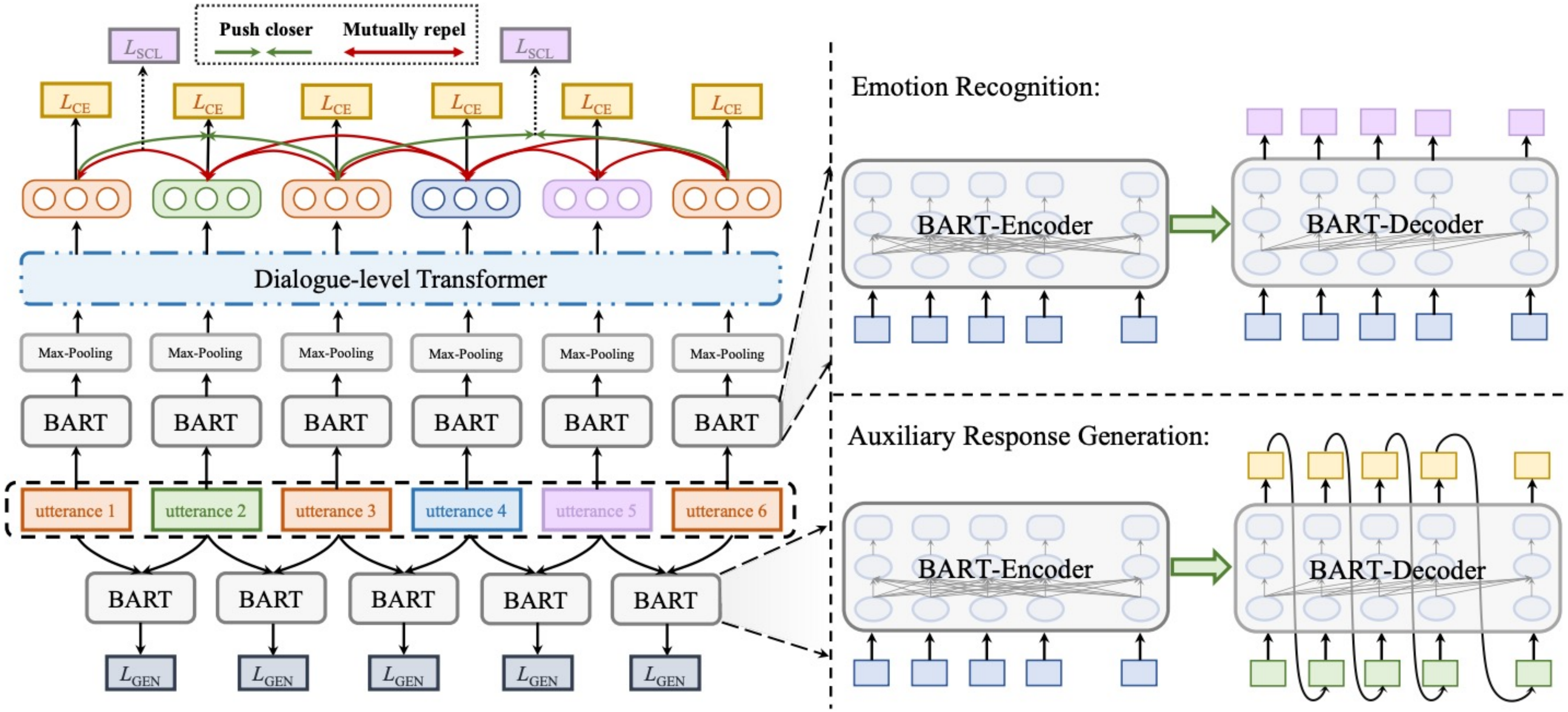
$$\mathcal{L}_{\text{Gen}} = -\sum_{i=1}^{N} \log p(u_{t+1}|u_t, \boldsymbol{\theta}),$$



Method

- Modeling Training
  - The total loss is the weighted sum of cross entropy loss, contrastive loss and generation loss.

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



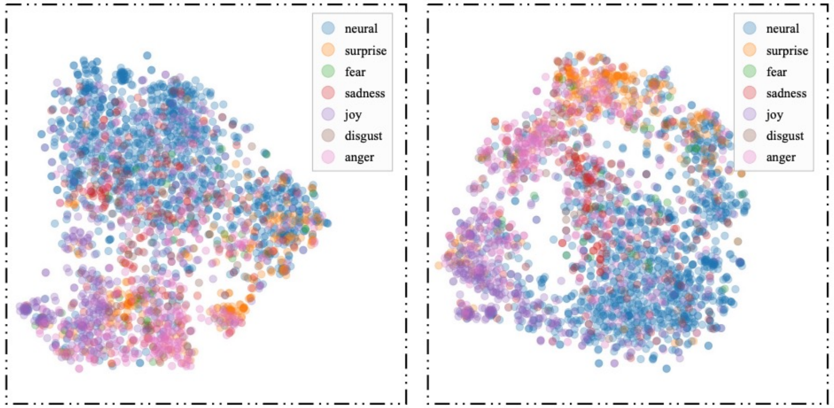
Experiments

- Metrics:
  - MELD, EmoryNLP and IEMOCAP: weighted average F1
  - DailyDialog: micro-F1 (ignore the label “neutral” when calculating the results)

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
TODKAT	68.23	64.75	43.12	42.68	61.33	61.11	52.56	58.47
BART	66.89	64.28	47.53	39.01	53.76	53.31	54.57	53.16
CoG-BART	**69.70 (±0.31)**	63.66 (±0.63)	**49.08 (±0.50)**	37.57 (±0.76)	**67.00 (±0.47)**	**64.10** (±0.51)	**56.46 (±1.03)**	54.71 (±0.99)

- Compared with pre-train-based models, our method yields very competitive results in four datasets.

Experiments



The t-SNE visualization when α is 0 and 0.8

Metric	Weighted Average F1					
Datasets	α=0.2	α=0.4	α=0.6	α=0.8	β=0.1	β=0.2
MELD	68.71	**69.70**	69.23	69.17	69.70	69.53
IEMOCAP	65.86	65.75	65.02	**67.00**	65.38	65.57
EmoryNLP	48.82	**49.08**	48.32	47.74	49.08	48.73

Quantitative Analysis

- Qualitatively, supervised contrastive learning clarifies the boundaries of the different emotional categories.

- The proportion of supervised contrastive loss in achieving optimal results varies between datasets.

Experiments

Dataset	MELD	IEMOCAP
Methods	Weight-Avg-F1	
CoG-BART	69.70	67.00
-Gen	68.50 (↓1.20)	66.63 (↓0.37)
-SCL loss	67.71 (↓1.99)	62.01 (↓4.99)
-Speaker	68.28 (↓1.42)	56.68 (↓10.32)
-Gen, SCL loss	67.32 (↓2.38)	64.68 (↓2.32)
-SCL loss, Speaker	67.05 (↓**2.65**)	54.57 (↓**12.43**)
-Gen, Speaker	68.39 (↓1.31)	56.14 (↓10.86)
-Dialog-Trans	69.40 (↓0.30)	66.83 (↓0.17)

Ablation Study

- Supervised contrastive loss is of great help for this task
- Speaker information plays a vital role in IEMOCAP

Thanks