ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096909

PAGE: A POSITION-AWARE GRAPH-BASED MODEL FOR EMOTION CAUSE ENTAILMENT IN CONVERSATION

Xiaojie Gu[1], *Renze Lou*[2], *Lin Sun*[1*], *Shangxin Li*[1]

[1]Department of Computer Science, Hangzhou City University, Hangzhou, China
[2] Department of Computer Science and Engineering, Pennsylvania State University, State College, USA

ABSTRACT

Conversational **C**ausal **E**motion **E**ntailment ($\mathrm{C_2E_2}$) is a task that aims at recognizing the causes corresponding to a target emotion in a conversation. The order of utterances in the conversation affects the causal inference. However, most current position encoding strategies ignore the order relation among utterances and speakers. To address the issue, we devise a novel position-aware graph to encode the entire conversation, fully modeling causal relations among utterances. The comprehensive experiments show that our method consistently achieves state-of-the-art performance on two challenging test sets, proving the effectiveness of our model. Our source code is available on Github[1].

***Index Terms*—** Emotion cause entailment, graph neural networks, position encoding

1. INTRODUCTION

For a target utterance transmitting a specific emotion, the $\mathrm{C_2E_2}$ aims to identify the causal utterances from the conversation history responsible for the target emotion. This novel task is essential to design current dialogue agents, such as empathetic response [1] and emotion counseling [2]. Besides, it provides a potential way to improve the interpretability of affect-based models [3].

Lee et al. [4] first proposed emotion cause extraction (ECE), who pointed out the importance of this task and considered it in a sentence-level classification paradigm. Early research employed rule-based strategies [5] and traditional machine learning approaches [6, 7] to deal with this problem. As a further step toward the ECE, $\mathrm{C_2E_2}$ considers a more challenging conversation scenario. So far, only a few studies [3, 8] have put the finger on $\mathrm{C_2E_2}$. Porial et al. [3] solely paired one target utterance with other utterances, which loses contextual information and breaks the causal relationship between utterances. Li et al. [8] utilized commonsense knowledge to facilitate causes recognization. From our insight view, the contextual information has a significant effect on utterance understanding, and it is difficult to determine whether they have a corresponding causality without considering the position between them [9, 10]. For that, a naive way is to directly concatenate the absolute position embedding with the utterance representation [11, 12]. However, this scheme constantly leads to the aggregation of uninformative context, bringing inference noise (i.e., the causal-irrelevant context). To this end, Xia et al. [13] utilized a multi-head attention mechanism to weigh the position information. Ding et al. [9] set the window size to consider only the adjacency of target utterances. While these position encoding strategies target the documents rather than conversations. Therefore, they neglected the vital inter-speaker dependency, which is essential in understanding conversation [14].

To address the above issues, we propose PAGE (**P**osition-**A**ware **G**raph-based model for **E**motion cause entailment), in which we devise a novel relative position encoding schema to distinguish utterances of different speakers for better reasoning. Intuitively, relative position plays a vital role in conversation-based causal inference. For example, there are two causal-relevant utterances from distinct speakers, namely "*Hey, you wanna see a movie tomorrow?*" and "*Sounds like a good plan.*". If the order is reversed, we probably do not realize that "a good plan" refers to "see a movie". Therefore, we construct the position relationship based on the relative distance between the different utterances of the inter-speaker. Furthermore, there is an explicit topological relationship between the emotion and the cause [15], so we leverage graph neural networks to encode the entire conversation context. We evaluate our approach on the latest benchmark dataset. The experimental results show that our method gains competitive performance compared with other strong baselines. In summary, the main contributions of this paper are as follows:

• We propose a position encoding strategy that can enhance emotion cause entailment and the understanding of conversation context in $\mathrm{C_2E_2}$. We design a novel position-aware graph to better aggregate the entire conversation.

• We conduct comprehensive experiments to demonstrate the effectiveness of our approach and provide a thorough analysis. Our model achieves state-of-the-art performance on the

*Corresponding author. This work was supported by the Natural Science Foundation of Zhejiang Province (No. LGN22F020002) , the Key Research and Development Program of Zhejiang Province (Grant No. 2022C03037) and the School-level Innovation and Entrepreneurship Training Program for College Students (No. X202001027).

[1]`https://github.com/XiaojieGu/PAGE`

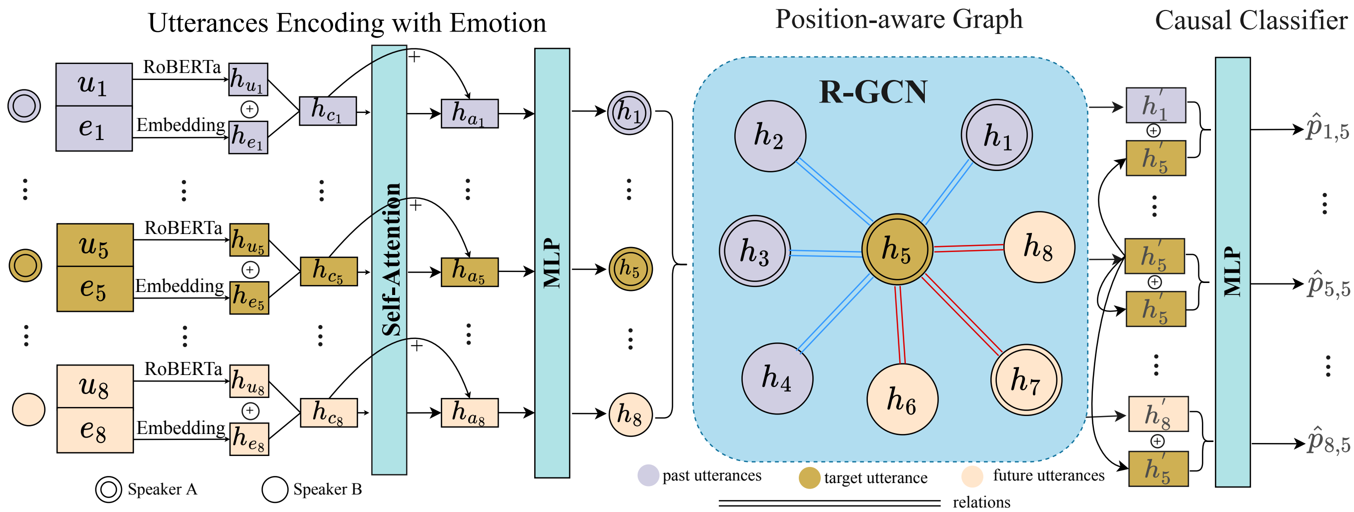

Fig. 1. Overall architecture of the PAGE.

benchmark dataset, especially improving 8.1% absolute in F1 score on the IE test set.

2. METHODOLOGY

As shown in Figure 1, our framework consists of three components, namely Utterances Encoding with Emotion (§ 2.1), Position-aware Graph (PaG) (§ 2.2) and Causal Classifier (§ 2.3). Given a conversation $C = \{u_1, u_2, \ldots, u_k\}$ where there are target utterances $u_t \in C$ of speaker $S_t \in \{A, B\}$. Each u_t is labeled with a specific non-neutral emotion (e.g., *"happiness"*, *"anger"*, *" sadness"*). Our goal is to identify the causal utterances in the conversation history of u_t. That is, whether a u_i ($i \leq t$) is a cause for u_t or not.

2.1. Utterances Encoding With Emotion

For an utterance $u_n = \{w_1, w_2, \ldots, w_m\}$ consisting of m words, we add two special tokens $[CLS]$ and $[SEP]$, at the beginning and end of it, respectively. Then, we use Roberta [16] to conduct sentence-level encoding and take the hidden state of the last layer as word-level representation:

$$h_w = \text{RoBERTa}\left([CLS], w_1, w_2, \ldots, w_m, [SEP]\right), \quad (1)$$

where $h_w \in \mathbb{R}^{m \times d_R}$ and $d_R = 768$ denotes the dimension of word-level representation in RoBERTa. Then, we get utterance representation $h_u \in \mathbb{R}^{d_u}$ by a linear projection $W_u \in \mathbb{R}^{d_R \times d_u}$ on h_w.

Emotion information in the conversation can pass among speakers [17], which is beneficial for entailment [18]; thus, we concatenate emotion embedding with the utterance representation, i.e., $h_c = h_e \oplus h_u$. To capture the utterance features from multiple aspects, we employ a multi-head self-attention mechanism [19], which contributes to sentence-level sentiment analysis [20]. The value of the Q, K, and V vectors are the same as h_c:

$$head_N = \text{softmax}\left(\frac{QK^T}{\sqrt{d_u}}\right)V, \quad (2)$$

where $head_N \in \mathbb{R}^{\frac{d_u}{N}}$ and N is the number of head. We concatenate heads together to get the attention output h_a and add it to h_c by $x = h_a + h_c$. Then, followed by multilayer perceptron (MLP) with a single hidden layer and a residual connection:

$$h_n = \sigma\left(MLP(x)\right) + x, \quad (3)$$

where the output dimension of the MLP is $\mathbb{R}^{d_u}$ and $\sigma(\cdot)$ is a sigmoid function.

2.2. Position-aware Graph

To alleviate long-term dependency between utterances in long conversations, we utilize graph neural networks to perform position-aware encoding. We design a directed graph denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$. The utterance $u_n \in \mathcal{V}$ in the conversation denotes a node, whose initial representation is h_n, and $r_{o,t} \in \mathcal{R}$ is the type of an edge $(u_o, r_{o,t}, u_t) \in \mathcal{E}$, where u_t denotes the target utterance and u_o ($1 \leq o \leq n$) denotes other utterances.

The relative position plays a significant role in the transformation of causal information between utterances; thus, we represent the type of edges between nodes in relative position relations. Moreover, the sequence relationship between the utterances of the same or different speakers facilitates the information understanding in the utterances and enhances entailment. For this, we construct a positional relation $r_{o,t}$ based

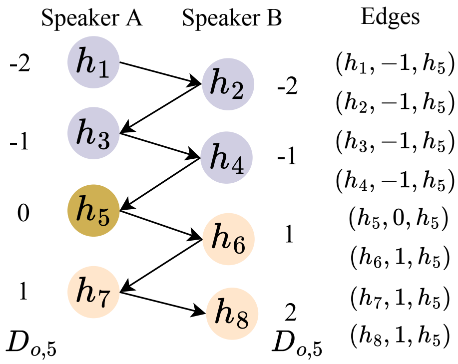

Fig. 2. An example of our relative position encoding. The target utterance is u_5 and window size $w = 1$. Purple represents past utterances, orange represents target utterance and cream represents future utterances.

on a relative distance $D_{o,t}$ that denotes the distance between the target utterance u_t and a neighboring utterance u_o:

$$D_{o,t} = \begin{cases} \frac{o-t}{2} & S_o = S_t \\ -1 & S_o \neq S_t \; and \; t = o \pm 1. \\ \frac{o-t-1}{2} & others \end{cases} \tag{4}$$

As the example shown in Figure 2, the target utterance u_5 has the same relative distance from both surrounding utterances u_3 and u_4 in the case of the same or different speakers, i.e., $D_{3,5} = D_{4,5}$, they are considered to be in the same positional relationship, i.e., $r_{3,5} = r_{4,5}$. For the utterance u_j where $j > t$, the utterances after the target utterance are not dominant compared to the parts before the target utterance [21], but they still can serve as additional context information to conversation understanding. We cannot completely ignore them, so we treat the parts after the target utterance as the same relative position relation. In this paper, we set it to 1. Furthermore, previous work indicates that most causes are located near the emotion [10], and if two utterances are far apart, then the causal relationship between them would diminish. We do not need to give them the same attention compared to those near the target utterance. To alleviate this issue, we set a window size in which the relative position relation between u_t and u_j is set to the same. In short, we get relative position relation $r_{o,t}$ from u_o to u_t by:

$$r_{o,t} = \begin{cases} -w & D_{o,t} < -w \\ D_{o,t} & D_{o,t} \geq -w \; and \; o \leq t. \\ 1 & o > t \end{cases} \tag{5}$$

We use R-GCN [22], which can integrate different relationships between nodes to get the final utterance representation which is beneficial to the position-aware transformation:

Test set	Conv.	Utt.	Avg.	Pos. Pairs	Neg. Pairs
DD	225	2,405	10	1,894	26,814
IE	16	665	41	1,080	11,305

Table 1. The statistics of RECCON test set, where ”DD” and ”IE” stands for the RECCON-DD and RECCON-IE test sets, respectively. Avg. represents the number of utterances per conversation on average.

$$h_t^{'} = \sigma \left(\sum_{r \in \mathcal{R}} \sum_{o \in \mathcal{N}_t^r} \frac{1}{c_{t,r}} W_r h_o + W_0 h_t \right), \tag{6}$$

where $\mathcal{N}_t^r$ denotes the set of neighboring nodes of node t under the relationship r, $c_{t,r}$ is a regularization constant, $W_r \in \mathbb{R}^{d_u \times d_u}$ is the trainable parameter to transform the neighborhood node o with relationship r.

2.3. Causal Classifier

We concatenate the final represent representations of the target utterance u_t and the other utterance u_o, then employ MLP with a single hidden layer and a sigmoid function to yield logits $\hat{p}_{o,t}$ and use cross-entropy loss function to do binary classification:

$$\hat{p}_{o,t} = \sigma \left(\text{MLP} \left(h_o^{'} \oplus h_t^{'} \right) \right). \tag{7}$$

3. EXPERIMENT

3.1. Experiment Setting

Dataset. We use a recently proposed benchmark dataset for $\mathrm{C_2E_2}$, namely RECCON [3], where each utterance in the conversations is attached with a human-annotated emotion label. It is sampled from two popular datasets of emotion recognition task in conversation [23, 24]. To comprehensively evaluate the model, there are two test sets in RECCON: RECCON-DD and RECCON-IE. It is worth noting that the data source of the RECCON-DD is the same as those in the training set, while RECCON-IE is not. In other words, the RECCON-IE stands for a demanding cross-domain evaluation. Specifically, as shown in Table 1, the average length of each conversation in IE is longer than in DD. Besides, the emotional shifts in IE conversations are more frequent than in DD, which requires a more complex emotional understanding capacity of the model [3]. Following Poria et al. [3], we omit those rare-appeared future causes and adopt the “Fold-1” of RECCON as the negative samples.

Implementation Details. For training, we choose the cross entropy loss function and set the learning rate to 3e-5 with a batch size of 4. The dimension d_u of utterance representations

Model	DD			IE		
	Neg. F1	Pos. F1	Macro F1	Neg. F1	Pos. F1	Macro F1
Base[3]	88.74	64.28	76.51	95.67	28.02	61.85
ECPE-MLL[25]	94.68	48.48	71.59	93.55	20.23	57.65
ECPE-2D[9]	94.96	55.50	75.23	**97.39**	28.67	63.03
RankCP[26]	**97.30**	33.00	65.15	92.24	15.12	54.75
KEC♣ [8]	95.74$_{(\pm 0.05)}$	66.76$_{(\pm 0.33)}$	81.25$_{(\pm 0.17)}$	86.08$_{(\pm 0.46)}$	19.72$_{(\pm 1.71)}$	52.9$_{(\pm 0.8)}$
PAGE	95.80$_{(\pm 0.06)}$	**68.80**$_{(\pm 0.11)}$	**82.30**$_{(\pm 0.05)}$	96.41$_{(\pm 0.25)}$	**45.96**$_{(\pm 0.82)}$	**71.19**$_{(\pm 0.52)}$
-w/o PaG	93.36$_{(\pm 0.46)}$	52.94$_{(\pm 0.97)}$	73.15$_{(\pm 0.31)}$	84.53$_{(\pm 2.0)}$	21.62$_{(\pm 0.32)}$	53.07$_{(\pm 0.89)}$

Table 2. The results on RECCON. We use Macro F1 as an overall metric, while the Pos. F1 and Neg. F1 represent F1 score on positive and negative pairs, respectively. ♣: since KEC [8] only report the results on DD, we run the KEC algorithm on the IE test set and report the results.

and hidden size of MLP are set to 300, and the number of attention heads and window size are set to 6 and 3, respectively. For the hyperparameter $c_{t,r}$, we fix it to 2.

3.2. Results

We compare our approach with solid baselines. As shown in Table 2, the scores of the first four baselines are reported by Poria et al. [3], which are the best-run results among several repeated experiments. For better comparison, we follow Li et al. [8] (i.e., KEC), reporting the average F1 score and the corresponding variance over five random runs.

Overall, we achieve state-of-the-art performance on both test sets under two main metrics (i.e., Pos. F1 and Macro F1), except the Neg. F1 due to the imbalance in the number of the Pos. and Neg. samples, as shown in Table 1. Compared to their methods, our relative position encoding strategy can fully consider inter-speaker dependency, which can effectively enhance the understanding of utterances. Since the conversations in the IE set are longer than those in the DD set, as shown in Table 1, the emotion cause shifts more in the IE set. Therefore, detecting the causes in the IE set is more challenging than in the DD set. Owing to our PaG structure, which has advantages in aggregating long-term contextual information, our model still maintains excellent results on IE, with a notable performance gap compared with other baselines. Since the conversations in real-world applications are mostly verbose [27], this promising performance demonstrates the robustness and practicality of our approach.

To further prove the effectiveness of our position encoding strategy, we conduct an additional ablation study. As shown in the bottom row of Table 2, after removing the PaG, the performance drops significantly, especially in those lengthy conversations of the IE test set. This result indicates that position is beneficial to emotion cause inference in the conversations, which is particularly helpful to those long conversations, demonstrating the outstanding generalization capacity and potential practicality of our model.

3.3. Effect of Window Size

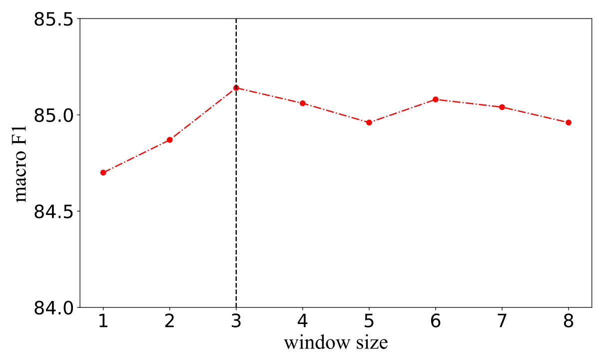

Fig. 3. Performance on the DD validation set with varying window size.

The window size may have an underlying effect on the model performance. Hence, we experiment to investigate the performance fluctuation from various window sizes. Figure 3 presents the results. Theoretically, a larger window size can lead to more position categories in the graph, but it increases the complexity of the graph network. As shown in Figure 3, when we increase the window size from 1 to 3, the performance is improved because the diverse position categories benefit the context understanding. However, the large window size (> 3) has adverse effects due to the inference noise from the long-distance utterances (i.e., causal-irrelevant context). Thus we choose window size = 3 in the PAGE.

4. CONCLUSION

In this paper, we propose a position-aware graph for C_2E_2 task. Our framework takes advantage of an effective position encoding strategy incorporating inter-speaker dependency, thus enhancing the capacity for complex emotion cause reasoning. Our PAGE model achieves SOTA performance on two challenging test sets, with 1.1% and 8.1% improvement at Macro F1 compared to previous models.

5. REFERENCES

[1] Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung, "Generating empathetic responses by looking ahead the user's sentiment," in *ICASSP*. IEEE, 2020, pp. 7989–7993.

[2] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang, "Towards emotional support dialog systems," *arXiv preprint arXiv:2106.01144*, 2021.

[3] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al., "Recognizing emotion cause in conversations," *Cognitive Computation*, vol. 13, no. 5, pp. 1317–1332, 2021.

[4] Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang, "A text-driven rule-based system for emotion cause detection," in *NAACL HLT 2010*, 2010, pp. 45–53.

[5] Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang, "Emotion cause detection with linguistic constructions," in *COLING*, 2010, pp. 179–187.

[6] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz, "Detecting emotion stimuli in emotion-bearing sentences," in *CICLing*. Springer, 2015, pp. 152–165.

[7] Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, Patricio Martínez-Barco, et al., "Emocause: an easy-adaptable approach to emotion cause contexts," in *ACL*, 2011.

[8] Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou, "Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge," *arXiv preprint arXiv:2205.00759*, 2022.

[9] Zixiang Ding, Rui Xia, and Jianfei Yu, "Ecpe-2d: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction," in *ACL*, 2020, pp. 3161–3170.

[10] Hanqi Yan, Lin Gui, Gabriele Pergola, and Yulan He, "Position bias mitigation: A knowledge-aware graph model for emotion cause extraction," in *ACL*, 2021, pp. 3364–3375.

[11] Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia, "From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification," in *AAAI*, 2019, pp. 6343–6350.

[12] Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang, "Context-aware emotion cause analysis with multi-attention-based neural network," *Knowledge-Based Systems*, vol. 174, pp. 205–218, 2019.

[13] Rui Xia, Mengran Zhang, and Zixiang Ding, "Rthn: A rnn-transformer hierarchical network for emotion cause extraction," *IJCAI*, 2019.

[14] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *EMNLP-IJCNLP*, 2019, pp. 154–164.

[15] Agnes Moors, "Automatic constructive appraisal as a candidate cause of emotion," *Emotion Review*, vol. 2, no. 2, pp. 139–156, 2010.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang, "Emotionflow: Capture the dialogue level emotion transitions," in *ICASSP*. IEEE, 2022, pp. 8542–8546.

[18] Guimin Hu, Guangming Lu, and Yi Zhao, "Bidirectional hierarchical attention networks based on document-level context for emotion cause extraction," in *EMNLP*, 2021, pp. 558–568.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Sheng-Chieh Lin, Wen-Yuh Su, Po-Chuan Chien, Ming-Feng Tsai, and Chuan-Ju Wang, "Self-attentive sentimental sentence embedding for sentiment analysis," in *ICASSP*. IEEE, 2020, pp. 1678–1682.

[21] Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang, "Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge," in *EMNLP*, 2021, pp. 1204–1214.

[22] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.

[23] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[24] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *IJCNLP (Volume 1: Long Papers)*, 2017, pp. 986–995.

[25] Zixiang Ding, Rui Xia, and Jianfei Yu, "End-to-end emotion-cause pair extraction based on sliding window multi-label learning," in *EMNLP*, 2020, pp. 3574–3583.

[26] Penghui Wei, Jiahao Zhao, and Wenji Mao, "Effective inter-clause modeling for end-to-end emotion-cause pair extraction," in *ACL*, 2020, pp. 3171–3181.

[27] Wenxiang Jiao, Michael Lyu, and Irwin King, "Real-time emotion recognition via attention gated hierarchical memory network," in *AAAI*, 2020, pp. 8002–8009.