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/ICASSP49357.2023.10094810

SDTN: SPEAKER DYNAMICS TRACKING NETWORK FOR EMOTION RECOGNITION IN CONVERSATION

Jiawei Chen, Peijie Huang, Guotai Huang, Qianer Li, Yuhong Xu*

College of Mathematics and Informatics, South China Agricultural University, China
{jw_chen, gthuang, li}@stu.scau.edu.cn, {pjhuang, xuyuhong}@scau.edu.cn

ABSTRACT

Emotion Recognition in Conversation (ERC) has considerable prospects due to its wide range of applications. Most existing works integrate speaker information statically and capture a relatively consistent atmosphere in conversation. However, these works poorly track the emotional state dynamics of each party in a conversation and focus on emotion consistency. The speakers' emotional states are independent but influence each other during the conversation. To address the above issues, we propose a **S**peaker **D**ynamics **T**racking **N**etwork (SDTN) for ERC. Specifically, SDTN can dynamically track the local and global speaker states during emotional flow in conversation and capture implicit stimulation of emotional shift. Extensive experiments on MELD and EmoryNLP datasets demonstrate the superiority and effectiveness of our proposed SDTN model, and confirm that every designed module consistently benefits the performance.

***Index Terms*—** Emotion Recognition, Dialogue System, Emotion Shift, Conversation

1. INTRODUCTION

Emotion Recognition in Conversation (ERC) is an important research topic due to its wide applications in many important tasks, such as empathetic dialogue generation [1], social media analysis [2], intelligent systems [3] and so on. The ERC task requires understanding how interlocutors express their emotions during conversations and classifying each utterance into a fixed set of emotion categories.

Unlike vanilla emotion recognition of the plain text, ERC is dynamic and highly correlated with its speaker and context information, especially for multi-turn conversations, which hold complex dependency between speakers. As shown in Figure 1, the emotional dynamics here depend on both the previous utterances and their associated speakers' emotional states. It has been argued that humans perceive emotions not only through the current utterance but also from its surrounding utterances [4]. Therefore, ERC models require a strong ability to model context dependencies and speaker relations and capture the dynamics of a specific speaker's emotion.

* Corresponding author.

Speaker	Utterance	Emotion Label
Phoebe	Coming through! Oh! Coming through! Oh! Hello! Hi! No! Right! Coming through!	fear
Monica	Oh well, it's not so bad.	neutral
Fireman	Yeah, most of the damage is pretty mostly contained in the bedrooms.	sadness
Phoebe	Oh!	surprise
Rachel	My God!	surprise

Fig. 1. A snippet of a dialog sample from the MELD Dataset.

Recent related works addressed contextual dependencies and speaker relations using numerous approaches. Basically, they can be divided into two categories [5]: static speaker-specific modeling [6, 7, 8, 9, 10] and dynamic speaker-specific modeling [11, 12, 5]. The former utilizes speaker information attached to each utterance to specify connections between utterances. The latter adopts intra- or inter-dependencies dynamically to facilitate modeling the current speaker state. Meanwhile, existing ERC models pay more attention to the speakers' emotion consistency while giving less consideration to the emotion shifts. However, the speakers' emotion changes dynamically because of the stimulation during a dialogue [3], and emotion recognition errors are more prone to occur when emotion shifts happen. Thus, the ability to track speaker dynamics, including the emotion shifts throughout a dialogue, synergizes with better emotion classification [13]. Further, capturing implicit emotional stimulation in speaker interaction contributes to effectively recognizing the emotion of utterances in conversation.

In this paper, we propose a novel speaker dynamics tracking network to recognize the utterance's emotion by sufficiently considering the influence of local and global interaction on speaker emotional dynamics, namely SDTN. Specifically, SDTN has two main modules, i.e., the speaker interaction tracker (SIT) and the emotion state decoder module. The SIT aims to track speaker interaction to capture the implicit stimulation for the interlocutor dynamically. Furthermore, the emotion state decoder considers the effects of emotion consistency and shift on decoding emotional state sequences. The experimental results demonstrate the superiority and effectiveness of our SDTN model and confirm the importance

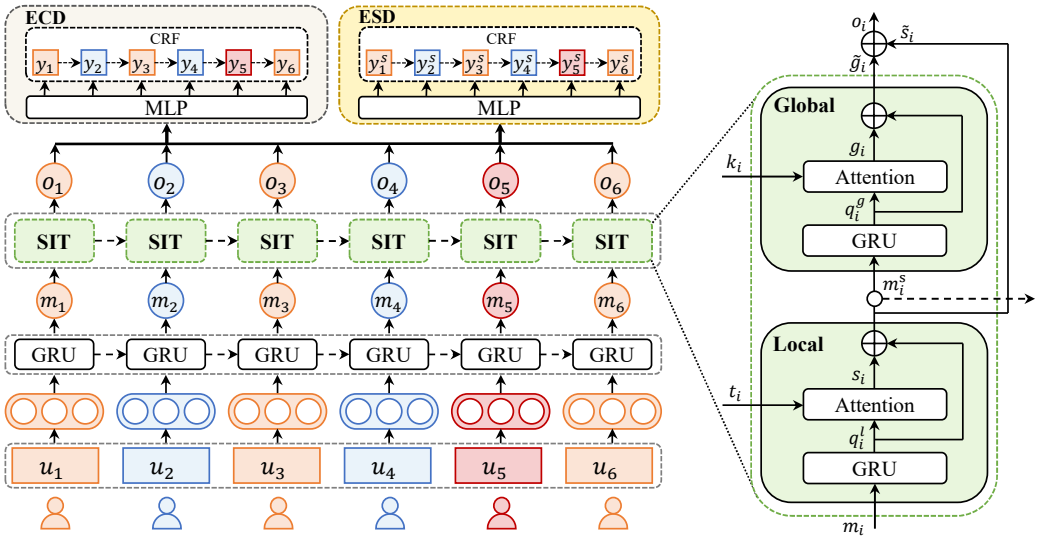

Fig. 2. The architecture of the proposed SDTN. We treat a whole conversation as input to our model.

of speaker dynamics tracking in ERC.

2. METHODOLOGY

In ERC task, a multi-party conversation with N consecutive utterances and corresponding speakers is denoted as $\{(u_1, p_1), (u_2, p_2), \ldots, (u_N, p_N)\}$. Each utterance u_i in conversation is spoken by a speaker p_i and has a discrete emotion label $y_i \in \mathcal{E}$, where $\mathcal{E}$ is the set of emotion labels. The ERC task aims to predict the emotion label y_t for a given u_t and p_t in conversation. In this section, our proposed SDTN for ERC is provided in Figure 2.

2.1. Textual Features

Following DAG-ERC [14], we fine-tune the pre-trained language model RoBERTa-Large [15] on each ERC dataset first and then freeze its parameters while training our SDTN. Each utterance is prepended with a special token [CLS] to obtain the input sequence. Afterward, we extract an utterance-level representation $\mathbf{u}_i$ from the [CLS]'s embedding of the last layer in RoBERTa with a dimension of 1024.

2.2. Context-level Modeling

Context-level Representation. Given the sequential nature of the conversation, we employ a vanilla unidirectional GRU [16] to capture the contextual relationship of utterances. Then we use context-level representation $\mathbf{c}_i$ form context memory representation $\mathbf{m}_i$ via a linear layer as follows:

$$\mathbf{c}_i, \mathbf{h}_i^c = \overrightarrow{\mathbf{GRU}}_{\mathcal{C}} \left(\mathbf{u}_i, \mathbf{h}_{i-1}^c\right), \tag{1}$$

$$\mathbf{m}_i = \mathbf{W}^q \, \mathbf{c}_i + \mathbf{b}^q, \tag{2}$$

where $\mathbf{W}^q$ and $\mathbf{b}^q$ are trainable parameters.

2.3. Speaker Interaction Tracker

The SIT aims to adequately model the speaker interaction and capture the implicit stimulation in the conversation, which consists of local and global interaction tracker modules.

Local Interaction Tracker. Since the speaker's emotional state is partly affected by the stimulation from surrounding interlocutors' utterances and the emotional legacy of the speaker's previous utterance, we keep track of the local interaction for each utterance in conversation to capture local speaker states.

Specifically, the context memory representation vector $\mathbf{m}_i$ is fed into GRU to learn the intrinsic logical order of local interactions in speakers' memory:

$$\mathbf{q}_i^l, \mathbf{h}_i^l = \overrightarrow{\mathbf{GRU}}_{\mathcal{L}} \left(\mathbf{m}_i, \mathbf{h}_{i-1}^l\right), \tag{3}$$

where $\mathbf{q}_i^l$ is the output vector of GRU. Then, we use $\mathbf{q}_i^l$ as the query and local memory $\mathbf{t}_i = \{\mathbf{m}_j \mid \forall j, \phi(u_i) \leq j \leq i\}$ as the key and value to implement the local attention mechanism, where $\phi(u_i)$ is the last previous utterance expressed by the same speaker of u_i, generating the local interaction state vector of $\mathbf{u}_i$ as follows:

$$\alpha_i^l = \mathbf{Softmax}\left(\mathbf{W}_l \left(\mathbf{q}_i^l \odot \mathbf{t}_i\right) + \mathbf{b}_l\right), \tag{4}$$

$$\mathbf{s}_i = \sum_{j=1}^{N} \alpha_i^l \mathbf{t}_i, \tag{5}$$

where $\mathbf{W}^l$ and $\mathbf{b}^l$ are trainable parameters, $\odot$ is an element-wise product operation. $\mathbf{s}_i$ is the the local interaction state of $\mathbf{u}_i$. Finally, we concatenate $\mathbf{q}_i^l$ and local interaction state $\mathbf{s}_i$ to obtain the final local speaker state $\tilde{\mathbf{s}}_i = \left[\mathbf{q}_i^l \| \mathbf{s}_i\right]$.

Global Interaction Tracker. We utilize the final local speaker state $\tilde{\mathbf{s}}_i$ of $\mathbf{u}_i$ to obtain the state memory repre-

sentation $\mathbf{m}_i^s = \mathbf{W}_s\tilde{\mathbf{s}}_i + \mathbf{b}_s$ via a linear layer, where $\mathbf{W}^s$ and $\mathbf{b}^s$ are trainable parameters.

Another GRU is used to learn the intrinsic logical order of local speaker states. Then, we employ $\mathbf{q}_i^g$ as the query and previous local speaker states $\mathbf{k}_i = \left\{\mathbf{m}_j^s \mid \forall j, j \leq i\right\}$ as the key and value to implement the global attention, generating the global interaction state vector $\mathbf{g}_i$ of $\mathbf{u}_i$ as follows:

$$\mathbf{q}_i^g, \mathbf{h}_i^g = \overrightarrow{\mathbf{GRU}}_{\mathcal{G}}\left(\mathbf{m}_i^s, \mathbf{h}_{i-1}^g\right), \tag{6}$$

$$\alpha_i^g = \mathbf{Softmax}\left(\mathbf{W}_g\left(\mathbf{q}_i^g \odot \mathbf{k}_i\right) + \mathbf{b}_g\right), \tag{7}$$

$$\mathbf{g}_i = \sum_{j=1}^{N} \alpha_i^g \mathbf{k}_i, \tag{8}$$

where $\mathbf{W}^g$ and $\mathbf{b}^g$ are trainable parameters, $\odot$ is an element-wise product operation. $\tilde{\mathbf{g}}_i = [\mathbf{q}_i^g \| \mathbf{g}_i]$ is the the final global speaker state of $\mathbf{u}_i$. Then, the final speaker state representation $\mathbf{o}_i = [\tilde{\mathbf{s}}_i \| \tilde{\mathbf{g}}_i]$ is the concatenation of the final local and global speaker state.

2.4. Emotion State Decoder

We employ two conditional random field (CRF) [17] layers to capture the sequential information of emotions and yield the final emotion labels of each utterance.

Emotion Consistency Decoder. The emotion consistency decoder (ECD) aims to capture the consistency of emotion labels in conversation. For an input set of speaker states $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_N\}$ and a sequence of emotion label predictions $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ of a conversation, the log-probability of the correct label sequence can write as follows:

$$\log(p(\mathbf{y} \mid \mathbf{O})) = s(\mathbf{O}, \mathbf{y}) - \log\left(\sum_{\tilde{\mathbf{y}} \in \mathbf{Y}} e^{s(\mathbf{O}, \tilde{\mathbf{y}})}\right), \tag{9}$$

where $s(\mathbf{O}, \mathbf{y})$ is the score for sequence $\mathbf{y}$ which is calculated by the trainable transition matrix $\mathbf{T}^c$ and the emission matrix $\mathbf{Q}^c$ from the final speaker states $\mathbf{O}$, following a linear layer and a softmax function. $\mathbf{Y}$ is the set of all possible label sequences. We optimize the CRF layer by maximizing the probability of ground truth emotion sequence p, the loss function is acquired:

$$L_{\mathrm{EC}} = -\sum_j \log(p(\mathbf{y} \mid \mathbf{O})), \tag{10}$$

Emotion Shift Decoder. Emotion shift describes the sudden changes in the same speaker's emotion, so the emotion shift information has a specific correlation with the main task. The emotion shift decoder (ESD) aims to capture the changes of emotion labels in conversation.

Following [18], we preprocess the datasets to obtain the emotion shift labels for the auxiliary task. We first select each utterance's preceding utterance of the same speaker. Then, if the emotion labels of the two utterances are the same, the emotion shift tag is set to 1; otherwise, it is set to 2. Additionally, we use 0 as the tag of the speaker's first utterance in the conversation. For a conversation, given a sequence of emotion shift tags $\mathbf{y}^s = \{y_1^s, y_2^s, \ldots, y_N^s\}$, the loss function of the CRF layer for emotion shift decoder is acquired:

$$\log(p(\mathbf{y}^s \mid \mathbf{O})) = s(\mathbf{O}, \mathbf{y}^s) - \log\left(\sum_{\tilde{\mathbf{y}}^s \in \mathbf{Y}^s} e^{s(\mathbf{O}, \tilde{\mathbf{y}}^s)}\right), \tag{11}$$

$$L_{\mathrm{ES}} = -\sum_j \log(p(\mathbf{y}^s \mid \mathbf{O})). \tag{12}$$

2.5. Model Training

During the training procedure, we treat the final representation $\mathbf{o}_i$ as input and employ the standard cross-entropy loss as objective function for speaker interaction tracker:

$$\mathbf{P}_i = \mathbf{Softmax}\left(\mathbf{W}_o\mathbf{o}_i + \mathbf{b}_o\right), \tag{13}$$

$$L_{\mathrm{CE}} = -\sum_{j=1}^{M}\sum_{i=1}^{N_j} y_{j,i} \log \mathbf{P}_{j,i}, \tag{14}$$

where $\mathbf{W}^o$ and $\mathbf{b}^o$ are trainable parameters. M is the number of dialogues in the train set, N_j indicates the number of utterances in the j-th dialogue. $y_{j,i}$ and $P_{j,i}$ denote the one-hot vector and probability vector for emotion labels of i-th utterance in the j-th dialogue, respectively. The SDTN are optimized via stochastic gradient descent during the training phase, and the total loss is the sum of losses from three components:

$$L = L_{\mathrm{CE}} + L_{\mathrm{EC}} + L_{\mathrm{ES}}. \tag{15}$$

3. EXPERIMENT

3.1. Datasets

We evaluate our proposed model on two benchmark ERC datasets. ***MELD*** [13] is a multi-modal multi-party conversation dataset collected from the TV series Friends. There are seven emotion labels: neutral, happiness, surprise, sadness, anger, disgust, and fear. ***EmoryNLP*** [19] is a multi-party conversation dataset collected from Friends, but varies from MELD in the choice of scenes and emotion labels. There are seven emotion types: neutral, sad, mad, scared, powerful, peaceful, and joyful. The statistics of them are shown in Table 1. We adopt weighted-average F1 and micro-F1 as the evaluation metrics.

3.2. Baselines

We compare the performance of the SDTN model with the following baselines:

Dataset	#Dial(Train/Val/Test)	#Utt(Train/Val/Test)
MELD	1,038/114/280	9,989/1,109/2,610
EmoryNLP	713/99/85	9,934/1,344/1,328

Table 1. The statistics of experimental datasets.

Model	MELD		EmoryNLP	
	W-Avg. F1	Micro F1	W-Avg. F1	Micro F1
RoBERTa	62.88	63.75	37.78	40.81
DialogueGCN	63.02	-	38.10	-
HiTrans	61.94	-	36.75	-
DialogXL	62.41	-	34.73	-
CoG-BART	64.81	65.95	39.04	42.58
DialogueRNN	63.61	-	37.44	-
DialogueCRN	63.42	-	38.91	-
SGED	63.34	-	38.47	-
COSMIC †	65.21	-	38.11	-
w/o KB	64.28	-	37.10	-
TODKAT †	65.47	-	**43.12**	42.68
w/o KB	63.97	-	33.79	-
Ours SDTN	**66.08**	**66.89**	39.48	**44.67**

Table 2. The overall performance of different pre-train-based baseline models on MELD and EmoryNLP. The models with † indicate that their results are not directly comparable with ours since they used external commonsense knowledge.

Static speaker-specific models: DialogueGCN [7], HiTrans [8], DialogXL [9], CoG-BART [10].
Dynamic speaker-specific models: DialogueRNN [11], DialogueCRN [12] and SGED [5].
Knowledge-enhanced models: Both COSMIC [20] and TODKAT [21] integrate external commonsense knowledge.

3.3. Implementation Details

We utilize the validation set to tune parameters on each dataset and adopt AdamW with a linear scheduled warm-up strategy. The parameters adjusted in this experiment include learning rate, dropout rate, and warm-up ratio. Specifically, the learning rate is 1e-5, except for the CRF layer, which is 1e-4. The results of our implemented models are all based on an average of 5 random runs on the test set.

3.4. Results and Analysis

Table 2 shows the main results of the proposed model and all compared baselines on the MELD and EmoryNLP datasets. Our proposed SDTN outperformed all baseline models except the TODKAT, which incorporates external commonsense knowledge on EmoryNLP dataset.

Compared with the static speaker-specific models, our SDTN outperforms CoG-BART by 1.27% on MELD and by 0.44% on EmoryNLP. For the dynamic speaker-specific models, our SDTN outperforms SGED by 2.74% on MELD and by 1.01% on EmoryNLP. Besides, compared with the models without external commonsense knowledge, our SDTN achieves significantly better performances on all datasets. For the knowledge-enhanced models, COSMIC and TODKAT can utilize external knowledge. It can be observed that compared with the models incorporating external knowledge, our SDTN still outperforms TODKAT by 0.61% on MELD in terms of the weighted-average F1 and by 1.99% on EmoryNLP in terms of the Micro-F1.

Model	MELD	
	W-Avg. F1	Micro F1
original SDTN	66.08	66.89
w/o SIT	64.69(↓1.39)	66.31(↓0.58)
w/o ECD + ESD	64.78(↓1.30)	66.03(↓0.86)
w/o ESD	65.29(↓0.79)	66.63(↓0.26)

Table 3. Ablation study on MELD dataset.

3.5. Ablation Study

Effects of speaker interaction tracker. From Table 3, it can be found that the performance of SDTN has a sharp decline with 1.39% when without speaker interaction tracker (SIT), which indicates the dynamic interaction information is vital for capturing speakers' emotional states.
Effects of emotion state decoder. ECD and ESD have a facilitating effect on modeling emotion dependence to some extent. To investigate the effects of the emotion state decoder in the SDTN, we perform the ablation studies by omitting the emotion consistency decoder (ECD) or emotion shift decoder (ESD). We can find that the SDTN without ECD and ESD or without ESD has a significant decline with 1.30% and 0.79%, respectively, which verifies the effectiveness of the decoders.

4. CONCLUSION

This paper proposes a Speaker Dynamics Tracking Network (SDTN) to fully track speakers' interaction and capture implicit stimulation in conversation to benefit understanding the speaker dynamics for the ERC task. In particular, we design the speaker interaction tracker to better track speaker interaction and capture implicit stimulation hierarchically. Then we apply two additional CRF layers to model speakers' emotion consistency and emotion shift during a conversation. The experimental results demonstrate the effectiveness of our proposed SDTN. Applying the SDTN to multi-modal emotion recognition in conversation will be our future work.

5. ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of Guangdong Province (2021A1515011864) and the National Natural Science Foundation of China (71472068).

6. REFERENCES

[1] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc. AAAI*, 2018, pp. 730–739.

[2] Runnan Li, Zhiyong Wu, Jia Jia, Yaohua Bu, Sheng Zhao, and Helen Meng, "Towards discriminative representation learning for speech emotion recognition," in *Proc. IJCAI*, 2019, pp. 5060–5066.

[3] Jiawen Deng and Fuji Ren, "A survey of textual emotion recognition and its challenges," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[4] James J Gross and Lisa Feldman Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion review*, vol. 3, no. 1, pp. 8–16, 2011.

[5] Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu, "Speaker-guided encoder-decoder framework for emotion recognition in conversation," in *Proc. IJCAI*, 2022, pp. 4051–4057.

[6] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou, "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations.," in *Proc. IJCAI*, 2019, pp. 5415–5421.

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

[8] Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu, "Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations," in *Proc. COLING*, 2020, pp. 4190–4200.

[9] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie, "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition," in *Proc. AAAI*, 2021, vol. 35, pp. 13789–13797.

[10] Shimin Li, Hang Yan, and Xipeng Qiu, "Contrast and generation make bart a good dialogue emotion recognizer," in *Proc. AAAI*, 2022, vol. 36, pp. 11002–11010.

[11] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "Dialoguernn: an attentive rnn for emotion detection in conversations," in *Proc. AAAI*, 2019, pp. 6818–6825.

[12] Dou Hu, Lingwei Wei, and Xiaoyong Huai, "Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations," in *Proc. ACL/IJCNLP*, 2021, pp. 7042–7052.

[13] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. ACL*, 2019, pp. 527–536.

[14] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan, "Directed acyclic graph network for conversational emotion recognition," in *Proc. ACL/IJCNLP*, 2021, pp. 1551–1560.

[15] 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.

[16] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[17] Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel, "Hierarchical pretraining for sequence labelling in spoken dialog," in *Findings of the Association for Computational Linguistics: EMNLP*, 2020, pp. 2636–2648.

[18] Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao, "Emotion recognition in conversations with emotion shift detection based on multi-task learning," *Knowledge-Based Systems*, vol. 248, pp. 108861, 2022.

[19] Sayyed M Zahiri and Jinho D Choi, "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*, 2018, pp. 44–52.

[20] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," in *Findings of the Association for Computational Linguistics: EMNLP*, 2020, pp. 2470–2481.

[21] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He, "Topic-driven and knowledge-aware transformer for dialogue emotion detection," in *Proc. ACL/IJCNLP*, 2021, pp. 1571–1582.