Emotion Prediction in Conversation Based on Relationship Extraction

Liu Yingjian[1], Wang Xiaoping[1] and Lei Shanglin[1]

***Abstract*— With the development of human-computer interaction systems, emotion recognition in conversation (ERC) has attracted increasing interest in recent years. Motivated by recent studies which have proven that generating emotional conversation responses can effectively improve the performance of the ERC model. However, accurate emotional response is complicated due to the limited number of dialogue samples. Therefore, we propose a simple and effective framework for emotion prediction in conversation based on relationship extraction (DiaRP), consisting of two curricula: (1) Dialogue Relationship Capture (DRC); and (2) Next Emotion Prediction (NEP). In DRC, we capture the current emotional self-dependence and interpersonal dependence according to the influence of self and others on the current moment emotion in the conversation. We integrate self-dependence and interpersonal dependence for NEP to predict their emotional state without current utterance. We also measure the similarity between recognized emotion distribution and predicted emotion distribution by the KL divergence. With the proposed model-agnostic DiaRP strategy, we observe a significant performance improvement over a wide range of existing ERC models and achieve new state-of-the-art results on three public ERC datasets.**

I. INTRODUCTION

Emotion recognition in conversation (ERC) is one of the most focused research fields in natural language processing (NLP), which aims to identify the emotion of each utterance in a conversation. With its potential applications in many fields, such as opinion mining in social media [1], empathic dialogue system construction [2, 3], and smart home systems [4, 5], this task has recently attracted the attention of a considerable number of NLP researchers.

Emotion is often reflected in social interpersonal communication, and analyzing the emotion of a single utterance without context may lead to ambiguity [6]. Therefore, ERC combined with the following information in history significantly contributes to emotion recognition, especially when the semantics of the utterance itself is ambiguous.

Recent studies have proved that generating emotional conversation responses can effectively improve the performance of the ERC model [7, 8]. This view can also be confirmed in real life: if we can understand the context well, we can easily guess the interlocutor's speech at the next moment according to historical clues. However, the small sample size of the ERC dataset results in poor quality of the generated conversation responses. Furthermore, if other semantically rich dialogue datasets are introduced for pre-training, a new problem of inconsistent data distribution is introduced.

[1]Liu Yingjian is with the School of Artificial Intelligence and Automation and the Key Laboratory of Image Processing and Intelligent Control of Education Ministry of China, Huazhong University of Science and Technology, Wuhan 430074, China. M202072868@hust.edu.cn

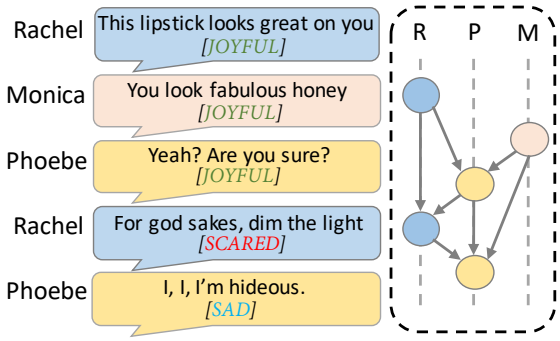

Fig. 1. In the relationship between different participants in the dialogue.

The problem arises from using limited contextual historical information to predict future information. There are direct speech-directed relations in actual conversations with a more explicit flow of affective tendencies. In contrast, this implicit information is challenging to give in the ERC task, and few studies have explicitly cared about this. In addition, the conditions under which the model generates semantically informative raw utterances are demanding and inconsistent with our essential emotion prediction purpose.

This paper solves the problem by designing a DiaRP module for the ERC task. DiaRP consists of Dialogue Relationship Capture (DRC) and Next Emotion Prediction (NEP). In DRC, we build the relationship between dialogue participants by explicitly making the emotion flow graph in the dialogue scenario. Expressly, the traditional ERC model often assumes that the emotional impact of the two participants in the previous round of dialogue is the greatest, which is reasonable in the two-person dataset. However, the model cannot obtain an effective dialogue relationship after extending it to the multi-part conversation scene. Therefore, DRC adopts the improved Attention mechanism, based on the relationship perception feature transformation of the information collected by the speaker's identity, and captures the utterance dependencies within and between remote speakers in the conversation from two subspaces. The NEP module is used to solve the emotional prediction problem under the lack of utterance in the next moment. Since the essential purpose of emotional dialogue generation is still emotional recognition, we only generate the future emotional state for recognition. Considering the lack of utterance in the next moment, we further relax the conditions and only take the emotional distribution identified by the traditional ERC task as the training goal. KL divergence is used as the optimization objective to measure the distance between the

two distributions.

Our DiaRP module is model-agnostic. We evaluate our approach on five representative ERC models and the results demonstrate that the proposed DiaRP leads to significant performance improvements.

In summary, our main contributions are as follows:

- We propose a DRC module to model the missing speaker-interlocutor relationship in the dialogue. In a multi-part conversation, capturing this relationship is more challenging than in a two-person conversation.
- We propose the NEP task to enhance context awareness. It implements the basic idea that minimizing the KL divergence between the recognized emotional distribution and the predicted emotional distribution by explicitly combining emotional self-dependence and interpersonal dependence can improve the context information perception ability of the ERC model.
- We conduct experiments on three ERC benchmark datasets. The empirical results show that our next emotion prediction task can effectively improve the overall performance of various ERC models, including the state-of-the-art.

II. RELATED WORKS

A. Emotion Recognition in Conversation

Unlike traditional emotion recognition which treats emotion as a static state, ERC considers emotion dynamic and flows between speaker interactions. Hazarika et al. [9] proposed a LSTM-based model to enable current utterances to capture contextual information in historical conversations. Jiao et al. [10] proposed a hierarchical GRU to effectively address the difficulty of capturing long-distance contextual information. DialogueRNN [2] modeled emotions dynamically based on the current speaker, contextual content, and emotional state by distinguishing specific speakers. By building directed graphical structures over the input utterance sequences with speaker information, DialogueGCN [11] applied a graph convolution network to construct inter- and intra-dependencies among distant utterances. Zhong et al. [12] proposed Knowledge-Enriched Transformer, which dynamically exploited external commonsense knowledge through hierarchical self-attention and context-aware graph attention. COSMIC [13] combined different commonsense knowledge and learned the interaction between the interlocutors in the dialogue. Wang et al. [14] proposed a relational graph attention network to encode the tree structure for sentiment prediction. DialogXL [15] modified the memory block in XLNet to store longer historical contexts and conversation-aware self-attention to handle multi-party structures. DAG-ERC [16] treated the internal structure of dialogue as a directed acyclic graph, which intuitively models the way information flows between long and short-distance contexts. Considering that utterances with similar semantics may have distinctive emotions under different contexts, CoG-BART [17] adopted supervised contrastive learning to enhance the model's ability to handle context information. GAR-Net [18] was an end-to-end graph attention reasoning network that took both word-level and utterance-level context into concern, aiming to emphasize the importance of contextual reasoning. Li et al. [19] designed an end-to-end ERC model called Emo-Caps, which could extract multi-modal information and the emotional tendency of the utterance effectively.

B. Dialogue Relationship Extraction

The relationship extraction (RE) task aims to identify relationships between entity pairs that exist in a document. Such tasks often require well-formed datasets, logically coherent, and explicit semantics. However, in the dialogue scenario, the simple logic of the utterance, the fuzziness of semantic reference, and more long-distance context dependencies make it more challenging to extract dialogue relations.

To solve this problem, DialogRE [20] proposed the first human annotation-based dialogue relationship extraction (DiaRE) dataset, DialogRE, which aims to support the relationship between the two arguments that arise in predictive conversations. Chen [7] proposed a DiaRE method based on a graphical attention network in which meaningful graphs connecting speakers, entities, entity types and corpus nodes are constructed to model the relationships between critical speakers. Sun [8] propose a utterance-aware graph neural network (ERMC-DisGCN) for ERMC, which design a relational convolution to lever the self-speaker dependency of interlocutors to propagate contextual information.

Unlike the previous researches, we predict the current emotion rather than recognize it to measure the context's influence on the current utterance and then capture the corresponding dialogue relationship. As far as we know, we are the first method to introduce DiaRE into ERC tasks.

III. PROPOSED FRAMEWORK

A. Task Definition

The ERC task is to recognize the corresponding emotion label e_t for each utterance $u_t^{p_i}$ in a sequence $\{u_1^{p_1}, u_2^{p_2}, \cdots, u_T^{p_m}\}$ from a conversation C containing m participants, where $p_i \in \{p_1, p_2, \cdots, p_m\}$ is the corresponding speaker identity, T represents the length of conversation. Moreover, two functions $s(\cdot)$ and $o(\cdot)$ are defined in conversation to output the last speech moment of the current speaker and the last speech moment of the corresponding interlocutor, respectively.

The recognition tasks in the above ERC can be summarized as $e_t = f_\theta(u_1, \cdots, u_t)$, where f_θ is ERC model. In order to better enhance the context awareness of ERC model, we define an emotion prediction task in conversation (EPC), that is, to predict the emotional state of t moment through the utterances before t moment, which is defined as $\hat{e}_t = g_\theta(u_1, \cdots, u_{t-1})$

B. Overview

Our framework consists of two parts: Dialogue Relationship Capture (DRC) and Next Emotion Prediction (NEP). In DRC, we extract the relationship in the conversation

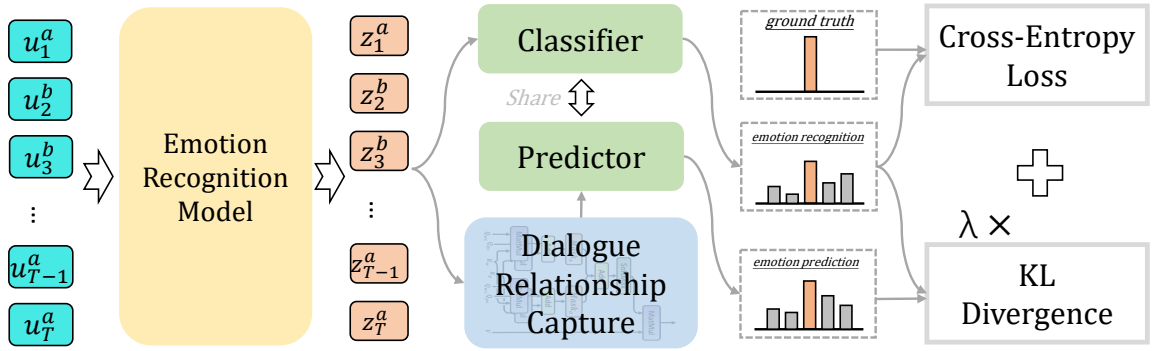

Fig. 2. The proposed DiaRP module for ERC. The DRC module captures self-dependence and interpersonal dependence based on the embeddings output of the ERC model to predict future emotions. The final optimization objective is determined by the classification loss and KL loss with λ weight.

according to the emotional influence of emotional self-dependence and interpersonal dependence through the attention mechanism. In the NEP module, we integrate self-dependence and interpersonal dependence to predict their emotional state without current utterances. We also measure the similarity between recognized emotion distribution and predicted emotion distribution by the KL divergence.

C. Dialogue Relationship Capture

In the ERC task, the speaker's emotional state at t moment is mainly affected by self-dependence and interpersonal dependence. Our EPC task also follows this assumption and captures such features through the Attention mechanism.

Specifically, concerning Self-attention's method of mapping query utterances and historical context to Q subspace and K subspace, respectively, we construct SO-Attention, which maps the query utterances to Q_s and Q_o subspaces, and historical context to the $K-V$ subspaces to calculate the speaker-specific contextual information and other participants' utterances influence in conversations. This method can selectively distinguish the contextual information within and between speakers, thus capturing the current speaker's self-dependence and interpersonal dependence. The specific formula is as follows:

$$\begin{aligned}\hat{v}_i &= \underbrace{\sum_{j<i, p_j=p_i} \frac{e^{q_{s,i}^T \cdot k_j}}{Z_i} v_j}_{v_i^s} + \underbrace{\sum_{j<i, p_j\neq p_i} \frac{e^{q_{o,i}^T \cdot k_j}}{Z_i} v_j}_{v_i^o}, \\ &= \sum_{j<i} Softmax\left(sim_{ij}\right) v_j\end{aligned} \tag{1}$$

where v^s and v^o represent self-dependence and interpersonal dependence respectively, Z_i is the normalized factor, and sim_{ij} represents correlation.

In addition, in the EPC task, the utterances in several rounds of close conversations tend to have a more significant impact. However, the traditional Transformer is not sensitive to the different effects of location. Therefore, we assume that the semantic contribution of historical statements in the dialogue to the current utterance shows a normal distribution and introduces Gaussian Self-attention proposed by Guo [21]. In addition, the mean and variance of Gaussian prior are challenging to quantify, so they are set as learnable parameters. The specific formula is as follows:

$$\begin{aligned}\hat{v}_i &= \sum_{j<i} \phi(d_{i,j}|\mu,\sigma) Softmax\left(sim_{ij}\right) v_j, \\ &= \sum_{j<i} Softmax\left(d_{ij}^{'} + sim_{ij}\right) v_j,\end{aligned} \tag{2}$$

where ϕ is a Gaussian distribution, μ and σ are their corresponding learnable parameters, and d_{ij} represents distance. Because the Gaussian distribution belongs to the exponential family distribution, it can be merged with Softmax so that the class multiplication becomes cumulative and the d_{ij} becomes $d_{ij}^{'}$ accordingly.

D. Next Emotion Prediction

For the next emotion prediction task, we combine the self-dependence and interpersonal dependence of the current moment to obtain the probability distribution of the prediction emotion class.

The emotional state captured by the DRC module is mapped to the classification space, and the probability distribution of the recognized emotional class P^{pred} is obtained by Softmax.

$$\begin{aligned}P_t^{pred} &= Softmax\left(g_\theta(v_t^s, v_t^o) \cdot W\right), \\ P_t^{recog} &= Softmax\left([g_\theta(v_t^s, v_t^o); z_t] \cdot W\right),\end{aligned} \tag{3}$$

where $W \in R^{2*h_e \times h_c}$ and $g_\theta(\cdot,\cdot)$ represents the emotional state generation network, and z_t is the input utterance u_t encoded output after the ERC model. Combined with the probability distribution of emotion class P^{recog} obtained by ERC tasks, the optimization objective of NEP is to minimize the KL divergence between the recognition probability and the prediction probability.

$$\begin{aligned}loss_{KL} &= D_{KL}(P^{recog}||P^{pred}), \\ &= \sum_{i=1}^{n} P_i^{recog} \cdot \log\left(\frac{P_i^{recog}}{P_i^{pred}}\right),\end{aligned} \tag{4}$$

where n represents the number of all utterances.

We use the shared weight matrix W to map the predicted emotion P^{pred} and the recognized emotion P^{recog}. This

is because Softmax is essentially a Proxy-based contrastive learning method, and the mapping matrix W contains the central distribution for each category. By sharing weights, our next emotion prediction task enables the ERC model to learn the category center embedding that can better distinguish each type of emotion.

IV. EXPERIMENTAL SETTINGS

A. Datasets

We evaluate our method on three published ERC datasets. Except for IEMOCAP, there is category imbalance in all datasets, especially DailyDialog. Following previous works (Ghosal et al. 2019; Zhong, Wang, and Miao 2019; Ishiwatari et al. 2020), we choose weighted-average F1 to validate IEMOCAP, micro-averaged F1 excluding the majority class (`neutral`) for DailyDialog, and weighted-average F1 for MELD and EmoryNLP. The detailed statistics of the datasets are reported in Table I.

IEMOCAP [22] is a dataset recorded as two-way conversational video clips, containing five sessions, where each session contains dyadic dialogues between two participators. Each dialogue is further segmented into utterances that are annotated with one of six emotion labels.

DailyDialog [23] is a multi-turn dialogue dataset, consisting of human-written daily communications. The language in DailyDialog dataset is human-written. Each utterance has a emotion label from one of seven categories.

EmoryNLP [24] is also a multi-dialogue dataset collected from $Friends$ TV script, which comprises 97 episodes, 897 scenes, and 12,606 utterances, where each utterance is annotated with one of the seven emotions.

B. Baselines

Since DiaRP is a model-agnostic framework, we choose the following five ERC models to verify whether DiaRP is able to further improve the performance of these models.

DialogueRNN [11] modeled emotions dynamically based on the current speaker, contextual content, and emotional state by distinguishing specific speakers.

COSMIC [13] trained commonsense conversational model COMET to extract commonsense features, and extracted contextual features from pre-trained transformer language model to model internal state, external state and intent state.

SKAIG [25] was enhanced by action information inferred from the past context and the intention implied by the future context. Meanwhile, it used CSK to represent rich edges with knowledge, and implied a graphics converter to process them.

DAG-ERC [16] treated the internal structure of dialogue as a directed acyclic graph to encode utterance, thus a more intuitive way to model the flow of information between long-distance conversation background and nearby context.

TODKAT [16] designed a topic-augmented language model (LM) with an additional layer specialized for topic detection and combined commonsense statements derived from a knowledge base based on the dialogue context.

C. Implementation Details

In this subsection, we mainly describe the implementation details of our proposed method. All of the baseline models mentioned above have released their source codes. We keep the same settings reported in the original papers during our experiments. For DiaRP, The hyperparameters include the weight of KL loss (Note that we only use the single-layer Attention mechanism to capture the conversation relationship). The results reported in our experiments are based on the average score of 5 random runs on the test set. A server with one NVIDIA 4090 GPU and Intel(R) Xeon(R) Silver 4210R CPU is used to conduct our experiments.

V. RESULTS AND ANALYSIS

A. Overall Results

Table II reports the overall experimental results, where "X+DiaRP" indicates training the model X with the proposed DiaRP framework. We can see that DiaRP improves the performance of all baseline models, showing the robustness and universality of our approach.

Overall, the performance improvement of DiaRP on ERC models with simpler feature extractors (i.e., COSMIC and SKAIG) is more significant. This is because the original ERC model inherently and implicitly can capture relationships in dialogues, while DiaRP gives it an explicit extraction framework and an optimization objectives. In contrast, DAG-ERC constructs a dialogue emotion flow graph based on a directed acyclic graph, which overlaps DRC module's functionality and results in limited performance improvement.

In addition, promoting the ERC model with the DiaRP module in the two-person dataset is higher than in the multi-part conversation dataset. This is because the ERP task predicts emotion through the context in the absence of current moment utterance, making the model need rich context information. However, in the EmotyNLP dataset, the dialogue length of a single sample is short and there are many participants in the dialogue, leading to emotional prediction more dependent on the utterance itself than the context.

B. Ablation Study

In order to prove the effectiveness of each module of DiaRP, we try to replace DRC and NEP modules on the DAG-ERC model. For the DRC module, we instead use the current speaker's interlocutor's speech from the previous round for the prediction of the current moment's emotion state, i.e., we replace self-dependence and interpersonal dependence with $u_{s(t)}$ and $u_{o(t)}$, respectively. For the NEP, we change the optimization objective of NEP to the ground truth, and the loss function is correspondingly changed to Cross-Entropy. The former is expressed as "X+Linear", while the latter is expressed as "X+CE".

As we can see in Table III, the presence of only two speakers in IEMOCAP even shows an increase in performance due to the clearer dialogue relationships, indicating that the dialogue relations in two-person dialogue are clear and rarely influenced by the distant utterances. For the multi-part dataset EmoryNLP, the number of speakers in each dialogue sample

TABLE I

THE STATISTICS OF DATASETS. *avg_utt* DENOTES THE AVERAGE NUMBER OF UTTERANCES IN A CONVERSATION.

Datasets	Conversations			Utterances			classes	type	avg_utt	Evaluation
	Train	Val	Test	Train	Val	Test				
IEMOCAP	108	12	31	5163	647	1623	6	two-person	47	Weighted-F1
DailyDialog	11118	1000	1000	87170	8069	7740	7	two-person	8	Micro-F1
EmoryNLP	713	99	85	9934	1344	1328	7	multi-party	11	Weighted-F1

TABLE II

THE OVERALL RESULTS ON THREE DATASETS. THE RESULTS OF BASELINE METHODS ARE FROM THE ORIGINAL PAPERS AND OUR REPRODUCTION.

Models	DailyDialog		IEMOCAP		EmoryNLP	
	Macro F1	Micro F1	W-Avg F1	Accuracy	W-Avg F1	Micro F1
DialogueRNN	55.95	41.80	62.75	63.40	-	-
COSMIC	58.48	51.05	65.28	66.37	38.11	37.89
SKAIG	59.31	51.82	65.79	66.15	37.57	37.13
DAG-ERC	59.33	-	68.03	68.61	39.02	38.94
TODKAT	58.47	-	61.33	62.35	43.12	43.68
DialogueRNN+DiaRP	56.27(↑ 0.32)	43.60	64.65(↑ 1.80)	65.50	-	-
COSMIC+DiaRP	59.11(↑ 0.63)	51.87	67.77(↑ **2.49**)	68.23	38.71(↑ 0.60)	38.86
SKAIG+DiaRP	59.83(↑ 0.52)	52.33	68.09(↑ 2.30)	68.27	38.50(↑ **0.93**)	38.26
DAG-ERC+DiaRP	59.73(↑ 0.40)	-	68.80(↑ 0.80)	68.74	39.46(↑ 0.44)	39.57
TODKAT+DiaRP	59.32(↑ **0.85**)	-	63.45(↑ 2.12)	64.54	43.79(↑ 0.67)	44.21

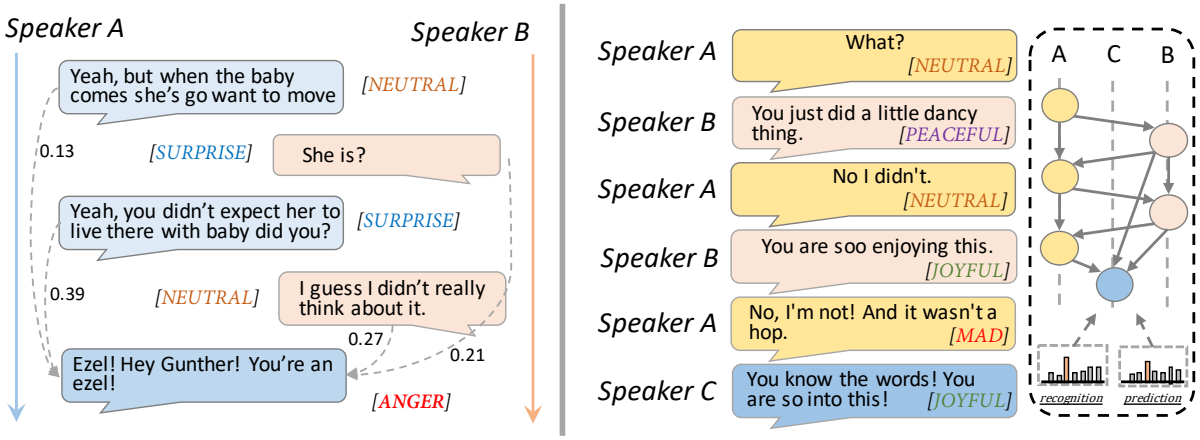

Fig. 3. Two conversation passages for case study. The ground truth emotion label of each utterance is given in the corresponding position, the boxes of different colors represent different speakers, and the final dark blue box represents the emotion to be predicted. (a) A case study of two-person conversation with the weight of dialogue relationship. (b) A case of emotion recognition and emotion prediction based on multi-part dialogue relationship.

TABLE III

ABLATION STUDY OF DIARP IN DAG-ERC

METHOD	IEMOCAP	EmoryNLP
DAG-ERC	68.03	39.02
DAG-ERC+DiaRP	68.80(↑ 0.77)	39.46(↑ 0.44)
DAG-ERC+Linear	69.02(↑ 0.99)	39.09(↑ 0.07)
DAG-ERC+CE	67.98(↓ 0.05)	39.27(↑ 0.25)

is different, so there is not necessarily a significant dialogue relationship in the neighbor moments, which reduces its emotion prediction ability. Meanwhile, the SO-Attention in DRC contains learnable Gaussian prior, which to some extent, enhances the influence of the latest round of dialogue, indicating the versatility of our method. The effect of Cross-Entropy loss is also worse, indicating that ground truth is not as rich in information as the distribution obtained by emotion recognition. Specifically, negative performance occurs in IEMOCAP using "X+CE", suggesting that accurate prediction of future emotion is contrary to the original ERC task; while "X+CE" still shows some improvement on the EmoryNLP dataset, demonstrating the need for the DRC module to model relationships on multi-part datasets

C. Case Study

Figure 3(a) shows a segment of a two-person conversation. It can be seen that the distribution of predictive emotion tends

to be consistent with that of identifying the emotion, and the attention weight also proves that it can effectively capture the relationship in long-distance conversation. Because the dialogue relationship in the multi-part dataset in Figure 3(b) is unclear, many baseline methods easily mistake the emotion for MAD when switching emotions. Most of our "X+DiaRP" methods can correctly identify the tone of this utterance, which shows that DRC can enhance the ability of relationship extraction in dialogue through NEP tasks.

VI. CONCLUSIONS

In this paper, we propose a simple and effective next emotion prediction task to improve the ability of the ERC model to capture context dependencies. DiaRP is a flexible module independent of the original ERC model, which can predict the future emotional state by capturing and fusing the influence relationship between different speakers in the dialogue. In the training process, DiaRP first captures the relationship between the speaker and interlocutor with the help of single-layer SO-Attention. Then, KL divergence is used to measure the similarity between recognized emotion distribution and predicted emotion distribution. Experiments on three benchmark datasets have proved the versatility and effectiveness of DiaRP. In the future, we plan to improve our approach in two ways. First, we will introduce commonsense information to compensate for the lack of context information in multi-part conversation datasets. Secondly, we try to present a more effective structure to integrate self-dependence and interpersonal dependence to generate the next moment's emotional state.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62236005, 61876209 and 61936004. (Corresponding author: Xiaoping Wang; e-mail: wangxiaoping@hust.edu.cn)

REFERENCES

[1] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, 2017.

[2] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[3] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "Moel: Mixture of empathetic listeners," *arXiv preprint arXiv:1908.07687*, 2019.

[4] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207*, 2018.

[5] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to-end dialogue systems with commonsense knowledge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[6] H. Zhou, M. Huang, and T. Zhang, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] Y. Sun, N. Yu, and G. Fu, "A discourse-aware graph neural network for emotion recognition in multi-party conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2949–2958.

[7] H. Chen, P. Hong, W. Han, N. Majumder, and S. Poria, "Dialogue relation extraction with document-level heterogeneous graph attention networks," *arXiv preprint arXiv:2009.05092*, 2020.

[9] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2594–2604.

[10] W. Jiao, H. Yang, I. King, and M. R. Lyu, "Higru: Hierarchical gated recurrent units for utterance-level emotion recognition," 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)*, 2019, pp. 397–406.

[11] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "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)*, 2019, pp. 154–164.

[12] P. Zhong, D. Wang, and C. Miao, "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)*, 2019, pp. 165–176.

[13] Ghosal, M. Deepanway, G. Navonil, and Alexander, "Cosmic: Commonsense knowledge for emotion identification in conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2470–2481.

[14] K. Wang, W. Shen, Y. Yang, X. Quan, and R. Wang, "Relational graph attention network for aspect-based sentiment analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3229–3238.

[15] W. Shen, J. Chen, X. Quan, and Z. Xie, "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 789–13 797.

[16] W. Shen, S. Wu, Y. Yang, and X. Quan, "Directed acyclic graph network for conversational emotion recognition," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1551–1560. [Online]. Available: https://aclanthology.org/2021.acl-long.123

[17] S. Li, H. Yan, and X. Qiu, "Contrast and generation make bart a good dialogue emotion recognizer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 002–11 010.

[18] H. Xu, Z. Yuan, K. Zhao, Y. Xu, J. Zou, and K. Gao, "Gar-net: A graph attention reasoning network for conversation understanding," *Knowledge-Based Systems*, vol. 240, p. 108055, 2022.

[19] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "Emocaps: Emotion capsule based model for conversational emotion recognition," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1610–1618.

[20] D. Yu, K. Sun, C. Cardie, and D. Yu, "Dialogue-based relation extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4927–4940.

[21] M. Guo, Y. Zhang, and T. Liu, "Gaussian transformer: a lightweight approach for natural language inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6489–6496.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[23] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, 2017, pp. 986–995.

[24] S. M. Zahiri and J. D. Choi, "Emotion detection on tv show transcripts with sequence-based convolutional neural networks," in *AAAI Workshops*, 2017, pp. 44–52.

[25] J. Li, Z. Lin, P. Fu, and W. Wang, "Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 1204–1214.