INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Dimensional Emotion Prediction based on Interactive Context in Conversation

Xiaohan Shi[1], *Sixia Li*[1], *Jianwu Dang*[1,2]

[1]Japan Advanced Institute of Science and Technology
[2]Tianjin key lab of cognitive computing and application, Tianjin University, China
s1810103@jaist.ac.jp, lisixia@jaist.ac.jp, jdang@jaist.ac.jp

Abstract

Emotion prediction in conversation is important for humans to conduct a fluent conversation, which is an underexplored research topic in the affective computing area. In previous studies, predicting the coming emotion only considered the context information from one single speaker. However, there are two sides of the speaker and listener in interlocutors, and their emotions are influenced by one another during the conversation. For this reason, we propose a dimensional emotion prediction model based on interactive information in conversation from both interlocutors. We investigate the effects of interactive information in four conversation situations on emotion prediction, in which emotional tendencies of interlocutors are consistent or inconsistent in both valence and arousal. The results showed that the proposed method performance better by considering the interactive context information than the ones considering one single side alone. The prediction result is affected by the conversation situations. In the situation interlocutors have consistent emotional tendency in valence and inconsistent tendency in arousal, the prediction performance of valence is the best. In the situation that interlocutors' emotional tendency is inconsistent in both valence and arousal, the prediction performance of arousal is the best.

Index Terms: affective computing, emotion prediction, Concordance Correlation Coefficient

1. Introduction

Human conversation is for exchanging ideas and communicating feelings. For the influent conversation, it is important for interlocutors to understand their emotions and predict the tendency of emotion's variation. Humans can express and understand the emotions of each other, while it is difficult for chatbots to understand the emotional changes and tendencies of the interlocutor in the conversation to continue the conversation more fluently. Therefore, predicting emotion variation is important for conducting a fluent human-machine conversation.

Emotion prediction in conversation is the task of predicting the coming emotional state of the speaker from previous context information. This task is still underexplored in the area of affective computing [1]. Among the studies on emotion prediction, Noroozi et al. [2] manually concatenated four speech turns as time series context information to predict coming emotional state categories with annotating of boredom, fear, joy, and sadness. Shahriar et al. [1] built a model to predict coming emotions with historical information from a single speaker, and predict the emotion categories: anger, happiness, neutral, and sadness. However, emotion in conversation is not an isolated state. It is an evolutionary state affected by the interaction of both interlocutors in the real dynamic interaction scenario [3]. For this reason, it is better to consider interactive context information on predicting emotion changes instead of using one single speaker's information alone. In addition, the influence of the interactive context information on emotion prediction may also be related to different conversation situations such as comforting, convincing, and so on. However, these problems are not explored and discussed in previous studies.

To address these problems, we propose a dimensional emotion prediction model to investigate the effects of the interactive context information on emotion prediction and the effects in different conversation situations. To investigate the effect of interactive context information, we compare prediction performances using context information of both interlocutors and one single speaker alone. To investigate the effects of different conversation situations on emotion prediction, we define the conversation situations and divide conversations into several groups. Then we compare the prediction performance in different groups. The context information deal within this study includes textual and acoustic information.

The rest of this paper is organized as follows. The proposed method is introduced in section 2. Section 3 gives an overview of the corpus and experimental setup. Section 4 describes the result and discussion. Finally, section 5 gives a conclusion of this paper.

2. Proposed Method

2.1. Emotion Prediction in Conversation

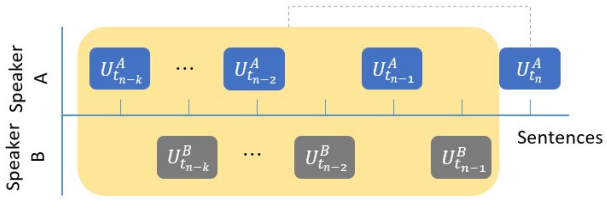

Figure 1: *Definition of context sequence in conversation*

Figure 1 shows the definition of context sequence in a conversion. There are two speakers in the given conversation, U^A reprensents the utterances of speaker A, U^B represents the utterances of speaker B, t is the number of the turns in the conversation.

When taking one speaker alone into account, for instance, speaker A, the context information for speaker A can be denoted as $(U^A_{t_{n-k}}, ..., U^A_{t_{n-2}}, U^A_{t_{n-1}})$. When taking both interlocutors into account, the interactive context information would become as $(U^A_{t_{n-k}}, U^B_{t_{n-k}}, ..., U^A_{t_{n-2}}, U^B_{t_{n-2}}, U^A_{t_{n-1}}, U^B_{t_{n-1}})$.

In this study, we use continuous values of Valence-Arousal dimension to describe the emotional state, thus the emotion label can be represented as $L_{t_n} = L_{Valence}, L_{Arousal}$, where L_{t_n} denotes the label of n-th turn. There are two common methods to annotate labels based on values in Valence-Arousal. One

http://dx.doi.org/10.21437/Interspeech.2020-1820

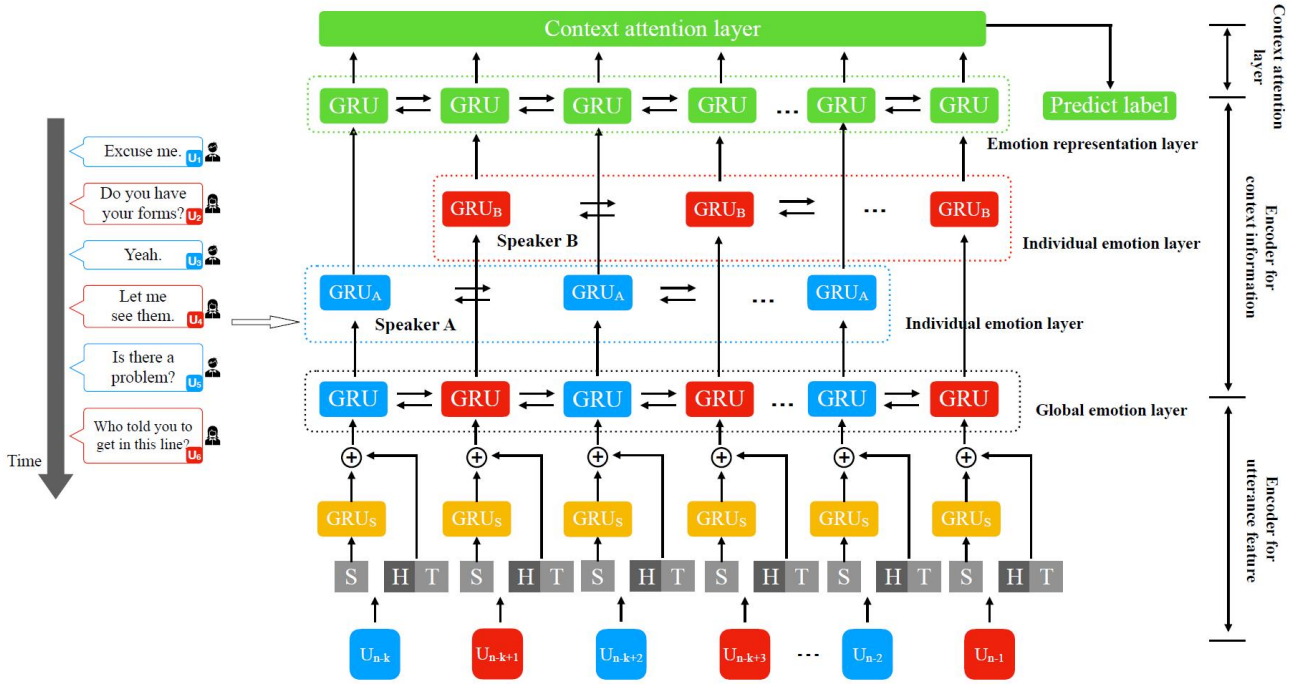

Figure 2: *The structure of emotion prediction model based on interactive context*

is using continuous values as labels which is usually for the regression analysis task [4]. The other one is converting values into discrete labels based on range of values which is usually for multi-classification task [5]. In this study, we follow the previous study [5] to convert continuous values into discrete labels based on value ranges in valence dimension and arousal dimension to annotate utterances.

2.2. Model Description

We propose a hierarchical model to predict the coming emotion using the interactive context in conversation. As shown in Figure 2, it contains three parts: encoder for the utterance-level feature, encoder for context-level information, and context attention layer. The encoder for utterance-level feature is used to concatenate acoustic feature and textual feature as utterance-level feature. The encoder for context-level information is set to a hierarchical structure, which is shown good performance on capturing interactive contextual information [6]. The context attention layer is used to output the emotion prediction and can be used to analyze the contribution of context information. Since each encoder deals with time series tasks and Gate Recurrent Unit (GRU) is suitable for capturing the temporal properties of the data, we use GRU for each encoder in our model.

The GRU network is an enhanced version of Recurrent Neural Network (RNN) that has the advantage of handling the vanishing gradient problem [7]. Each GRU cell consists of an update gate (z) and a reset gate (r) to control the flow of information. Let x be the input to the GRU network, W and U denote weight parameters and b denote the bias of GRU network. At time step t, the hidden state (h_t) can be computed as following equations:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{1}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{2}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(W_h x_t + U_t(h_{t-1} \odot r_t) + b_h) \tag{3}$$

In this structure, the spectrogram feature S from each utterance is encoded by the GRU_S into a vector representation on utterance level. Then, we concatenate spectrogram feature S, heuristic feature H [8], and textual feature T into one vector as utterance-level feature. To capture the interactive information of both interlocutors, the utterance-level feature is conveyed to the global emotion GRU layer of the encoder for context information. Then, based on the speakers that the utterance belongs to, the hidden state at each time step of the global emotion GRU layer is separately conveyed to the individual emotional layer GRU_A or GRU_B. The individual layer can capture the context information of the speakers. Finally, the emotion representation GRU layer and context attention layer are used to predict the coming emotional state. The attention weights from context attention layer can be used to analyze each utterance's contributions from interactive context information in the prediction.

For the preceding k turns as interactive context information, in the emotional expression GRU, we firstly calculate each hidden layer information of GRU to obtain X $(h_{t-k}, ..., h_t)$, then calculate the last hidden layer information of GRU, $q(h_t)$. Finally, we calculate the attention weight α_i of each utterance from interactive context information with dot product attention calculation. Then the result of the attention weight can be computed as:

$$\alpha_i = softmax(X_i^T * q) \tag{4}$$

$$att(q, X) = \sum_{i=1}^{k} \alpha_i X_i \tag{5}$$

Therefore, the attention weight from interactive context information $(\alpha_1, ..., \alpha_k)$ can represent the contributions of each utterance from interactive context information in the prediction.

2.3. Feature Extraction

In this study, we use the acoustic features and textual features to describe the interactive context information. Acoustic features consist of two parts: spectrogram feature and heuristic feature. We use the Mel filterbanks [9][10] to extract the spectrogram feature. To apply the spectrogram feature into our model, we divide spectrogram to a number of segments with a window of 5ms long and 3ms shift. Then, as mentioned above, we use the GRU_S shown in Figure 2 to encode segment-level features from spectrogram into utterance-level feature. For the heuristic feature, we use openSMILE [11] to obtain the heuristic features with the eGeMAPS feature set [12] for input utterance. For textual feature, we use BERT to obtain the representation for input text utterance, where pre-trained model BERT is showed good performance on representing semantic meanings in vector space [13].

3. Experiments

3.1. Corpus

Interactive Emotional dyadic Motion Capture (IEMOCAP) database is a widely used corpus in affective computing. It contains approximately 12 hours of audio-visual recordings [14]. The transcripts are also provided in the corpus. Each conversation in IEMOCAP has been segmented into utterances with continuous label in Valence-Arousal dimension and category label in 9 categories. There are two kinds of conversations in the corpus, improvise form and script form. To reduce the potentially confounding effect of semantic information disturbance, we only use the improvise form conversations.

To ensure the experiment process to be speaker-independent, we use 5-fold cross-validation for the experiment following the work of [15].

Based on the proposed method, we follow the previous study [16] to convert continuous values in valence dimension and arousal dimension into three classes based on the value ranges. Value from 1 to 2 is defined as label 'low class', value from 2 to 4 is defined as label 'neutral class', value from 4 to 5 is defined as label 'high class'.

3.2. Experiment Procedure

We conducted two experiments in this study. Experiment 1 is to investigate the effects of interactive context information on emotion prediction. Experiment 2 is to investigate the effects of interactive context information in different conversation situations.

In Experiment 1, we compare the performance in two experimental approaches with different context information. One considers context information from a single speaker, the other considers interactive context information from both interlocutors.

In Experiment 2, we define conversation situations by the Concordance Correlation Coefficient (CCC) score of both interlocutors' emotional tendencies. If CCC > 0 in both valence and arousal dimensions, it means that both interlocutors have similar emotional tendencies during the conversation in both valence and arousal dimensions. The conversation situation is labelled as POS-POS. If CCC < 0 in both valence and arousal dimensions, it means that both interlocutors have different emotional tendencies in valence and arousal dimensions. The conversation situation is labelled as NEG-NEG. If CCC < 0 in one of valence or arousal dimension and CCC > 0 in the other dimension, the conversation situation is labelled as POS-NEG and NEG-POS, respectively. Consequently, we define conversations into four situation groups as following:

Group 1 POS-POS $CCC_{Valence} > 0$ and $CCC_{Arousal} > 0$

Group 2 POS-NEG $CCC_{Valence} > 0$ and $CCC_{Arousal} < 0$

Group 3 NEG-POS $CCC_{Valence} < 0$ and $CCC_{Arousal} > 0$

Group 4 NEG-NEG $CCC_{Valence} < 0$ and $CCC_{Arousal} < 0$

We divide the corpus into the four groups, Table 1 shows the statistics of the conversation situation groups.

Table 1: *The statistics of each conversation situation groups*

	Group1	Group2	Group3	Group4
Conversations	42	26	8	4

In this study, we merged consecutive utterances in a turn into one utterance unit. Accordingly, the conversation can be represented in the form of a sequence of utterance units alternating by speakers as shown in Fig. 1. Since a merged utterance for one turn may contain multiple annotated utterances, the emotional label of merged utterance is computed as the average value of its constituent utterances. In the experiment, we use interactive context information of eight turns to predict the coming emotional state, in which four turns are from one speaker, respectively.

3.3. Experimental Setup

For the feature extraction, we use 40 Mel-filter banks to obtain the log Mel features and use openSMILE toolbox to obtain 88 dimensions heuristic features. The BERT pre-trained model used to extract textual feature is from bert-as-service [17], the output dimensions is 378. For the model, each GRU is set to one layer with 256 units and dropout rate 0.9. In the training process, the optimizer is set to Adam optimizer with learning rate of 0.00001.

4. Result and Evaluation

We use the Unweighted Average Recall (UAR) [18] as evaluation metric in our study.

Table 2 shows the results of the performance using the context information from one single speaker and from both interlocutors, respectively. One can see that using context information of two interlocutors achieved better performance than that using information from single speaker alone. The improvement is 7% for prediction of valence and 2% for prediction of arousal. Error reduction rate is 16.5% for the prediction of valence and 4.5% for prediction of arousal. It implies that it is better to consider both interlocutor's interactive context information for emotional prediction than only considering context information from single speaker alone.

Table 2: *Comparison of single speaker and two interlocutors*

	Valence	Arousal
Single speaker	60.61%	43.60%
Two interlocutors	67.12%	46.13%

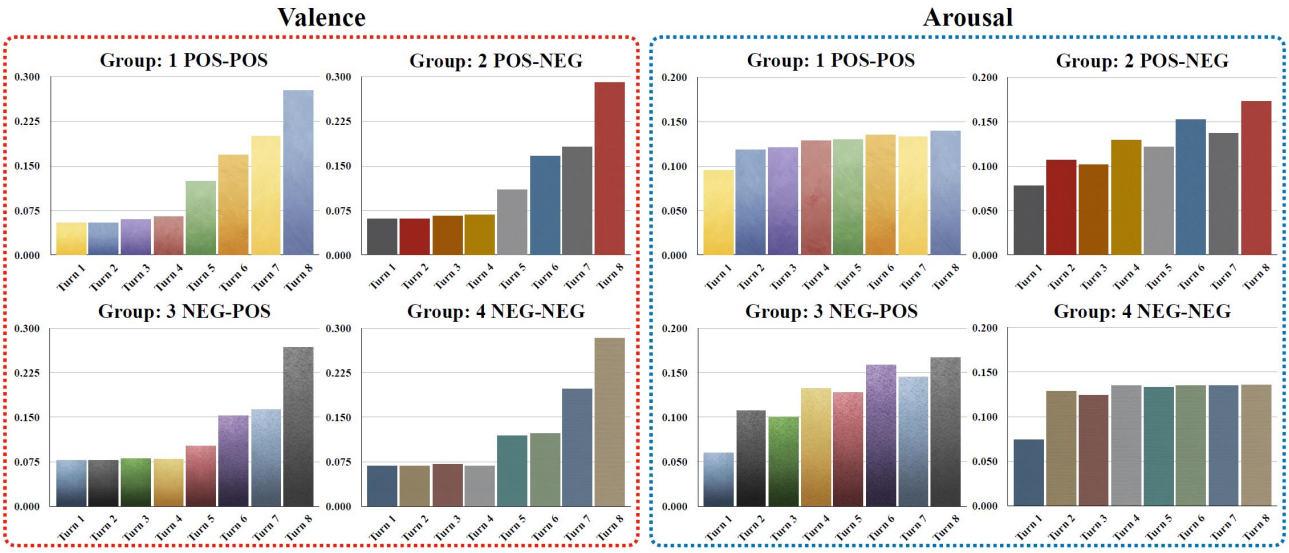

Figure 3: *Visualization of contributions of different turns of context information*

Table 3: *Comparison of each conversation situations using two interlocutor's interactive context information*

	Prediction of Valence	Prediction of Arousal
Pos-Pos	67.31%	40.69%
Pos-Neg	68.32%	46.58%
Neg-Pos	65.22%	41.48%
Neg-Neg	64.04%	54.88%

Table 3 shows the results of the performance in different conversation situations. One can see that the performance in different situations is different. In the situation that interlocutors have consistent tendency in valence while inconsistent tendency in arousal, the prediction performance of valence is the best. In the situation that emotional tendency of both interlocutors is inconsistent in both valence and arousal, the prediction performance of arousal is the best. In the database, we found that both interlocutors are dispute or appease in NEG-NEG conversation situations. In dispute with another speaker, both interlocutors are stuck on expressing their emotions, which is a special interaction in conversation. Such a kind of interactive information may make the prediction of arousal easier. Therefore, in NEG-NEG conversation situations, the performance of predicting arousal is better than the performance in other conversation situations.

To investigate the contributions of different components of interactive context information, we visualize the attention weights in different situations. Figure 3 shows the visualization results. For prediction of valence, the utterance closest to the nearest time step has the highest weight among all conversation situations, which means the interactive information closer to the last time step is more important. For the arousal prediction, in the POS-POS and NEG-NEG conversation situations, the utterance weights are almost the same among time steps, which indicates that the interactive information of both interlocutors is important. In conversation situations of POS-NEG and NEG-POS, the attention weights of the speaker's utterances in the preceding side are higher. It means the other side of the speaker's interactive information is more important.

5. Conclusions

In this paper, we proposed a dimensional emotion prediction model using interactive context information from both interlocutors in different conversation situations. The results show that the performance is better considering the interactive context information than only considering context information from a single side, and the performance is also related to the different conversation situations. In the situations, in which interlocutors have consistent emotional tendency in valence and inconsistent tendency in arousal, the prediction performance of valence was the best. In the situation that interlocutors' emotional tendency is inconsistent in both valence and arousal, the prediction performance of arousal was the best. It suggested that we should take interactive context information from both interlocutors in emotion prediction into account, and also consider different conversation situations such as comforting, convincing, where the emotion changes in some specific way.

6. Acknowledgements

This study is supported by JSPS KAKENHI Grant (20K11883), and partially by National Natural Science Foundation of China (No. 61876126).

7. References

[1] S. Shahriar and Y. Kim, "Audio-visual emotion forecasting: Characterizing and predicting future emotion using deep learning," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–7.

[2] F. Noroozi, N. Akrami, and G. Anbarjafari, "Speech-based emotion recognition and next reaction prediction," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2017, pp. 1–4.

[3] J. Zhao, S. Chen, J. Liang, and Q. Jin, "Speech emotion recognition in dyadic dialogues with attentive interaction modeling," *Proc. Interspeech 2019*, pp. 1671–1675, 2019.

[4] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion predic-

tion," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, 2012.

[5] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," *Proc. Interspeech 2019*, pp. 2578–2582, 2019.

[6] 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, 2019, pp. 6818–6825.

[7] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[8] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2666–2670.

[9] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[10] S. Lokesh and M. R. Devi, "Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method," *Cluster Computing*, vol. 22, no. 5, pp. 11 669–11 679, 2019.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] 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.

[15] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.

[16] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6705–6709.

[17] H. Xiao, "bert-as-service," https://github.com/hanxiao/bert-as-service, 2018.

[18] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.