EMOTIONFLOW: CAPTURE THE DIALOGUE LEVEL EMOTION TRANSITIONS

Xiaohui Song[⋆†‡], *Liangjun Zang*[†‡], *Rong Zhang*[⋆], *Songlin Hu*[†‡], *Longtao Huang*[⋆]

[†]Institute of Information Engineering, Chinese Academy of Sciences
[‡] School of Cyber Security, University of Chinese Academy of Sciences
[⋆]Alibaba Group
{songxiaohui,zangliangjun,husonglin}@iie.ac.cn
{kaiyang.hlt,stone.zr}@alibaba-inc.com

ABSTRACT

Emotion recognition in conversations (ERC) has attracted increasing interests in recent years, due to its wide range of applications, such as customer service analysis, healthcare consultation, etc. One key challenge of ERC is that users' emotions would change due to the impact of others' emotions. That is, the emotions within the conversation can spread among the communication participants. However, the spread impact of emotions in a conversation is rarely addressed in existing researches. To this end, we propose **EmotionFlow** for ERC with the consideration of the spread of participants' emotions during a conversation. EmotionFlow first encodes users' utterance by concatenating the context with an auxiliary question, which helps to learn user-specific features. Then, conditional random field is applied to capture the sequential information at emotional level. We conduct extensive experiments on a public dataset Multimodal EmotionLines Dataset (MELD), and the results demonstrate the effectiveness of our proposed model.

***Index Terms*—** Natural Language Processing, Dialogue System, Emotion Recognition

1. INTRODUCTION

With the rapid development of conversational AI research, emotion recognition in conversations (ERC) has received significant attention from NLP researchers[1, 2, 3, 4, 5, 6, 7]. Many efforts have been made to promote the performance and bring ERC to more application scenarios. ERC also plays an important role in a wide range of downstream tasks such as chat agents[1, 6], health-care applications([8]) and question answering[9], etc.

ERC aims at identifying different users' emotions at each turn in a conversation. Usually, the language used in a conversation is generally short and informative, so it is difficult to figure out emotions based on only one turn of textual utterance. As the example shown in Fig.1, a single word "*OK*" can express different emotions. When the previous user said "*Come on, she is gone, cheer up, ok?*", the response "*OK*" can be *sadness* or *frustrated*, but for the previous utterance "*Take me a cup of water*", the emotion of "*OK*" as the response is more likely to be *neutral*. This shows that appropriate context representation of users' utterances is essential for ERC. Numerous efforts have been devoted to the representation of the conversation context. Graph-based approaches[6, 1, 10, 2] build a graph upon the conversation, and complete the context modeling through the information aggregation on the graph. While recurrence-based approaches[11, 12, 4] usually encode the utterances turn by turn and model the flow of semantics in temporal sequence.

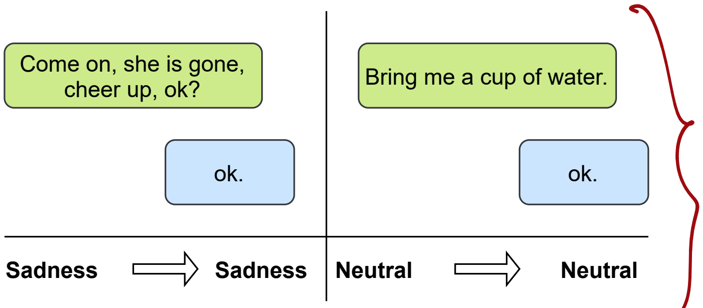
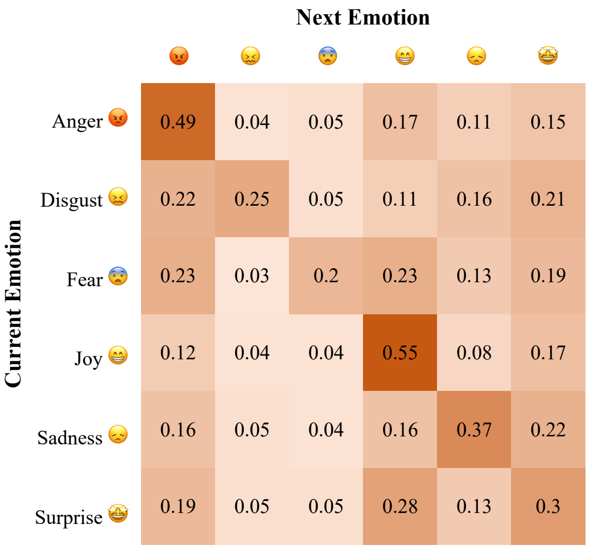


Fig. 1. An example of emotion spread in a conversation.

Existing ERC methods majorly make efforts to bridge semantics to emotions, but explicit modeling the spread effect of emotions in a conversation is rarely discussed. Intuitively, users' emotions would be affected by each other during the conversation such as the example in Fig.1. As shown in Fig.2, the emotions are not transferred randomly among the participants, and adjacent conversation turns are more likely to have the same emotions. Based on this observation, we propose a new method called **EmotionFlow** to capture the sequential information of emotions. EmotionFlow contains two sub-modules, an utterance encoder and a CRF layer. Inspired by the success of Transformers[13], we use the Transformer-based pre-trained language model RoBERTa[14] as the utterance encoder. Different from existing ERC methods, we construct the model input at each turn in a question answer-



Current \ Next	😡	😖	😨	😁	😞	🤩
Anger 😡	0.49	0.04	0.05	0.17	0.11	0.15
Disgust 😖	0.22	0.25	0.05	0.11	0.16	0.21
Fear 😨	0.23	0.03	0.2	0.23	0.13	0.19
Joy 😁	0.12	0.04	0.04	0.55	0.08	0.17
Sadness 😞	0.16	0.05	0.04	0.16	0.37	0.22
Surprise 🤩	0.19	0.05	0.05	0.28	0.13	0.3

Fig. 2. The transition probability between emotions of current turn and next turn, based on MELD training set statistics.

ing fashion. We concatenate the most recent k turns of utterances and corresponding user names as the passage, followed with an auxiliary question like "*how does [user_name] feel now?*". We add such questions at each turn to let the model learn user-specific features. After obtaining the user's probability distribution of emotions from the utterance encoder, we use it as emission scores and input to the CRF layer to learn the spread effect of emotions at dialogue level. We conduct extensive experiments on the widely used benchmark Multimodal EmotionLines Dataset(MELD)[15], and the results show that the proposed **EmotionFlow** achieves comparable performance with the state-of-the-art models.

2. RELATED WORKS

Traditional approaches to the ERC task can be divided into two categories according to their methods to model the conversation context. Recurrence-based models are committed to model the sequential information of the conversation. ICON[11] and HiGRU[16] both use the gated recurrent unit(GRU) to capture the context information. DialogRNN[12] is a recurrence-based method that models dialog dynamics with several RNNs. For those graph-based methods, DialogGCN[1] builds a graph upon the utterances nodes. RGAT[2] further improves the DialogGCN by adding the position embeddings. ConGCN[6] trades both speakers and utterances as nodes and builds a single graph upon the whole ERC dataset. DAG-ERC[5] uses a directed acyclic graph (DAG) to model the intrinsic structure within a conversation.

The state-of-the-art ERC methods are Transformers-based methods. DialogXL[3] improves pretrained transformers with enhanced memory and dialog-aware self-attention. KET[10] uses hierarchical Transformers with external knowledge. KAITML[7] builds a knowledge-aware incremental transformer with multi-task Learning, and it extracts commonsense knowledge from an external knowledge base. COSMIC[4] uses exploited ATOMIC[17] for the acquisition of `If-Then` commonsense knowledge. TODKAT[18] proposes a topic-augmented language model and it also utilizes the ATOMIC knowledge base.

3. METHODOLOGY

3.1. Problem Setup

Given a set of multi-party conversations $\mathcal{C}$, a collection of all speakers $\mathcal{S}$ and an emotion labels set $\mathcal{E}$, our goal is to identify speakers' emotions at each conversation turn. Each conversation contains a sequence of utterances $\{u_1, u_2, ..., u_N\}$ with corresponding speakers $\{s_1, s_2, ..., s_N, s_i \in \mathcal{S}\}$, where N is the number of turns. In this paper, we focus on the real-time settings of ERC task, in which the model takes only previous turns $\{(u_1, s_1), (u_2, s_2), ..., (u_t, s_t)\}$ as input and identifies the emotion of speaker s_t.

3.2. Semantic Context Modeling

We using the pretrained language model RoBERTa[14] to model the semantic context. As illustrated in Figure 3, we reformulate the emotion classification problem into a question answering task to let model learn speaker-specific features. We use the most recent k turns of utterances and corresponding speakers as passage, and construct a question at each turns like "How does [speaker] feel now?", the complete model input X_t at turn t is as follow:

$$X_t = [\texttt{<s>}, s_{t-k}, u_{t-k}, s_{t-k+1}, ..., s_t, u_t, \texttt{</s>}, Q] \quad (1)$$

where s_t, u_t are speaker and utterance of t-th turn. Q here is "How does s_t feel now?". `<s>` and `</s>` are special tokens for RoBERTa. Then we input X_t to the pretrained LM and use the `<s>`'s embeddings v_t at the last layer as the classification feature of utterance u_t.

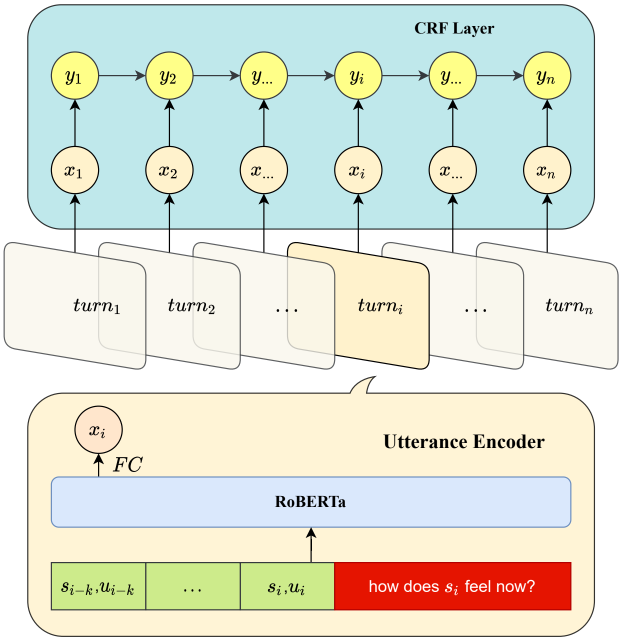
$$v_t = \text{RoBERTa}(X_t)[0] \quad (2)$$

We can obtain the pesudo-probability distribution $p_t \in \mathbb{R}^{|\mathcal{E}|}$ over all emotions of speaker s_t,

(Emotion Embedd. ?

$$p_t = W \cdot v_t + b \quad (3)$$

where $W \in \mathbb{R}^{|\mathcal{E}|*d}$ and $b \in \mathbb{R}^d$ are trainable parameters, d is the hidden size of RoBERTa and $|\mathcal{E}|$ is the size of emotion label set $\mathcal{E}$. During the training stage, we employ the standard cross-entropy loss as objective function for context modeling:



same order
+
a/q curr
↳ all turns

Fig. 3. The overview of EmotionFlow.

$$\mathcal{L}_1(\theta_1) = -\sum_{i=1}^{M}\sum_{t=1}^{N_i} \log p_{i,t} * y_{i,t} \tag{4}$$

where $M = |\mathcal{C}|$ is the number of conversations in the dataset, and N_i is the number of turns in the i-th conversation. $p_{i,t}$ is the probability over all emotion labels at turn t of i-th conversation, and $y_{i,t}$ is the corresponding ground truth label. θ_1 is the collection of trainable parameters.

3.3. Emotion Sequence Modeling

As shown in Figure 2, emotions during a conversation have obvious chronological correlation, in this paper, we use conditional random field(CRF) to capture the sequential information of emotions.

Given a conversation that contains N utterances, the goal of CRF layer is to maximize the probability of the ground truth emotion sequence over all possible emotion sequences:

$$\max(\mathrm{P}(y_1, y_2, .., y_N)|\boldsymbol{x}) \tag{5}$$

where $\boldsymbol{x} = [p_1, p_2, ..., p_N]$ are calculated via the utterance encoder, $p_i \in \mathbb{R}^{|\mathcal{E}|}$ is the probability distribution over all emotions at turn i and y_i are ground truth labels. According to the theory of linear chains CRF, P can write as follows:

$$\mathrm{P} = \frac{1}{Z(\boldsymbol{x})} \exp\left(p_{1,y_1} + \sum_{t=2}^{N}\left[g\left(y_{t-1}, y_t\right) + p_{t,y_t}\right]\right) \tag{6}$$

p_{i,y_i} here stands for the probability of j-th emotion at i-th turn calculated by the utterance encoder. $g \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ is a trainable parameter of CRF, which models the transition probability among emotions. $Z(\boldsymbol{x})$ is a normalization term and can be calculated via forward-backward algorithm [1]. In practice we compute the Z as follows:

$$Z_1 = p_1 \tag{7}$$

$$Z_t = \log \sum \left[\exp\left(Z_{t-1}^{\top} \oplus g \oplus p_t\right)\right] \tag{8}$$

$$Z(\boldsymbol{x}) = \sum_{k}^{|\mathcal{E}|} \exp(Z_{N,k}) \tag{9}$$

where $\oplus$ stands for the broadcast add[2] and N is the number of turns. We optimize the CRF layer via maximize the probability of ground truth emotion sequence P, loss function can write as follows:

$$\mathcal{L}_2(\theta_2) = -\log(\mathrm{P}) \tag{10}$$

where θ_2 is the collection of trainable parameters of CRF layer.

3.4. Training and Predicting

We treat a full conversation as an input. We first construct the QA-style input X_i for the utterance encoder at each turn, then calculate the emotion probability p_i for every X_i and obtain the classification loss $\mathcal{L}_1$. Finally, we feed $\boldsymbol{x} = [p_1, p_2, ..., p_N]$ into the CRF layer and get the negative log-likelihood loss $\mathcal{L}_2$. The utterance encoder and the CRF layer are optimized jointly via stochastic gradient descent during training phrase, and the total loss is the sum of losses from two sub-modules:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 \tag{11}$$

At the predicting stage, we run the Viterbi algorithm[3] over CRF layer to decode the best emotion sequence.

4. EXPERIMENTAL SETUP

4.1. Datasets

We conduct experiments on the Multimodal EmotionLines Dataset (MELD), it is a multimodal ERC dataset collected from the TV show *Friends*. There are 7 emotion labels including `neutral`, `joy`, `surprise`, `sadness`, `anger`, `disgust`, and `fear`. Details statistics of MELD are listed in Table 1.

The reason why we chose MELD to evaluate our method is the speakers' distribution. There are 274 speakers in

[1] https://en.wikipedia.org/wiki/Forward-backward_algorithm

[2] https://tvm.d2l.ai/chapter_common_operators/broadcast_add.html

[3] https://en.wikipedia.org/wiki/Viterbi_algorithm

Example: $[[p_{a,1}, p_{a,2}, p_{a,3}]$
$[p_{a,2}, - -]$
$[. . .]]$
matrix

	Train	Dev	Test	total
Conversations	1038	114	280	1432
Utterances	9989	1109	2610	13708
Speakers	260	47	100	274
Speakers >100	6	6	6	6

Table 1. Statistics of MELD.

Model	External Knowledge	Weighted F1
DialogueGCN	✗	58.10
RGAT	✗	60.91
HiTrans	✗	61.94
DialogXL	✗	62.41
DAG-ERC	✗	63.65
TODKAT w/o KB	✗	63.97
EmotionFlow(Ours)	✗	**65.05**
KAIMTL	✓	58.97
KET	✓	58.18
COSMIC	✓	65.21
TODKAT	✓	**65.47***

Table 2. Performance comparisons on the MELD testset. All the results are directly cited from their original papers. We divide the baseline models into two groups according to whether it uses external knowledge or not. Numbers on the bold font are the best performance of each group. (*The performance of TODKAT is updated by its authors in their official code repository.)

MELD, but only 6 protagonists have more than 100 utterances, which is helpful for the utterance encoder to learn speaker-specific features.

4.2. Implementation Details

For all experiments in this paper, we train the model on the training set of MELD for 10 epochs, select the best checkpoint using the develop set and then report weighted F1 scores on the test set. We concatenate the most recent 3 turns of utterances and the constructed speaker-specific question for each turn when constructing the model inputs. We use the base model of RoBERTa as our utterance encoder. As for other hyper-parameters, the batch size is 8, learning rates for the pretrained LM, the FC layer, and CRF layer are 1e-5, 1e-4, 1e-3, respectively. We conduct all experiments on a single NVIDIA V100 GPU. Due to space reasons, we will not discuss hyperparameter analysis here.

4.3. Results and Analysis

Table 2 shows the performance of different models on MELD testset. We can see that our model outperforms all the baselines without external knowledge, which shows the effectiveness of our proposed model for ERC. Meanwhile, we can find that state-of-the-art methods are those that incorporate external knowledge. Both COSMIC and TODKAT utilized the knowledge from ATOMIC[17], which is a commonsense knowledge base for `If-Then` reasoning. The way they use the knowledge is directly use a knowledge generation model called COMET[19]. COMET is a pretrained language model that finetuned on ATOMIC and ConceptNet. Comparing the performance of `TODKAT` and `TODKAT w/o KB` in Table 2, the COMET model can be used to explain the performance gap between EmotionFlow and TODKAT. In this paper, our focus is to capture the relationship of emotions in time series, so we did not incorporate external knowledge.

4.4. Ablation Study

To further demonstrate the effectiveness of EmotionFlow, we conduct ablation study on the MELD testset, as shown in Table 3. We directly use the output scores of the utterance encoder to predict emotions in the case of w/o CRF. EmotionFlow w/o QA in Table 3 means we delete the constructed question "How does speaker feel now?" for each turn. By doing this, the model needs to judge which sentence is the focus of this turn among the several sentences spliced together based on only the position embeddings of RoBERTa. The last row in Table 3 is the performance of a single RoBERTa. As expected, following Table 3, both CRF layer and QA-style input construction are essential to the strong performance of our model.

Model	Weighted F1
EmotionFlow	65.05
EmotionFlow w/o CRF	63.70
EmotionFlow w/o QA	63.55
EmotionFlow w/o [CRF,QA]	62.35

Table 3. Ablation study on MELD test set.

5. CONCLUSION

This paper proposes a new model called EmotionFlow for real-time emotion recognition in conversation (ERC). We apply an additional CRF layer to capture the emotion transition probability among speakers during a conversation. To better modeling the semantic context, we construct the input for each turn in a question answering fashion, by which the model can learn speaker-specific features from context. We conduct extensive experiments and the results show that the proposed EmotionFlow achieves comparable performance with the state-of-the-art methods. Moreover, through the ablation study, we confirmed the effectiveness of the CRF layer and QA-style input construction.

6. REFERENCES

[1] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.

[2] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7360–7370.

[3] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie, "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition," *arXiv preprint arXiv:2012.08695*, 2020.

[4] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," *arXiv preprint arXiv:2010.02795*, 2020.

[5] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan, "Directed acyclic graph network for conversational emotion recognition," *arXiv preprint arXiv:2105.12907*, 2021.

[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 *IJCAI*, 2019, pp. 5415–5421.

[7] Duzhen Zhang, Xiuyi Chen, Shuang Xu, and Bo Xu, "Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4429–4440.

[8] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[9] Mahmoud Azab, *Multimodal Character Representation for Visual Story Understanding*, Ph.D. thesis, University of Michigan, 2019.

[10] Peixiang Zhong, Di Wang, and Chunyan 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.

[11] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger 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.

[12] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6818–6825.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[14] 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.

[15] 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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[16] Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael 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.

[17] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 3027–3035.

[18] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He, "Topic-driven and knowledge-aware transformer for dialogue emotion detection," *arXiv preprint arXiv:2106.01071*, 2021.

[19] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi, "Comet: Commonsense transformers for automatic knowledge graph construction," in *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4762–4779.