ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413596

GENERATING EMPATHETIC RESPONSES BY INJECTING ANTICIPATED EMOTION

Yuhan Liu[1], *Jiachen Du*[1*], *Xiang Li*[1], *Ruifeng Xu*[1,2*]

[1]School of Computer Science, Harbin Institute of Technology (Shenzhen)
[2]PengCheng Laboratory
liuyuhan_hitsz@163.com, jacobvan199165@gmail.com, xiangli@stu.hit.edu.cn, xuruifeng@hit.edu.cn

ABSTRACT

Showing empathy and reacting to users' feeling are important social skills for current dialogue generation systems. In previous research, empathetic responses are generated by 1) only modeling the emotion of dialogue history or 2) indirectly leveraging the predicted emotion label of responses. In this paper, we propose a novel empathetic response generation method that incorporates the anticipated emotion into response generation by minimizing the divergence between distribution of responses' anticipated emotion and ground-truth emotion. The anticipated emotion is predicted by an auxiliary emotion predictor whose input is the previous utterances. Additionally, we treat the generation as deliberation process and design a two-round training method to refine the response iteratively. Experimental results show that the proposed model outperforms the previous state-of-the-art for emphatic dialogue generation task.

Index Terms— Natural Language Processing, Empathetic Dialogue, Anticipated Emotion Injection

1. INTRODUCTION

Improving the capability of chatbots in human-computer interaction and making them more user-friendly have become a significant problem in modern conversation systems. Directly infusing emotion into dialogue systems proves to be productive and instrumental [1, 2]. Since in our daily conversations, we may generate various feelings when facing different situations. Meanwhile, showing empathy is a prominent trait among human communications, which could elicit strong emotional resonance and deep understanding of personal experiences. Therefore, we focus on empathetic dialogue (EmpDialogue), which expects effectively modeling empathy for making chatbots more human-like [3].

In recent neural dialog systems, emotionally-aware conversation has gradually become a trend. As a sub-task, EmpDialogue has obtained much attention. And a key difference for EmpDialogue is that there are no assigned emotion labels for the next turn and chatbots can only look upon the emotion of the dialogue history in EmpDialogue. There are mainly two lines of work on EmpDialogue. The first line of work focus on generating response through understanding and modeling the emotion of the current state: MoEL [4] uses transformer [5] as encoders and decoders, then It applies different emotion decoders and softly combines outputs of them with a current emotion state distribution to get the final output; [6, 7] use pretrained language models [8] with simple multi-task method to get useful response. Instead, the second [9] leverage the reinforcement learning framework and treat predicted response emotion as the reward, then fuse this empathetic emotion signal into response generation while training.

Both previous cases have succeeded in generating empathetic responses, but have neglect some points of this task. The first ignore the excepted emotion of response when given the dialogue history and make it hard to explicitly show emphatic to the users. The second still cannot explicitly model response emotion feature during inference phase.

Therefore, to address the above issues, we build two independent emotion classifiers on the top of the generator such as GPT2 [10, 8]. One is called primary emotion predictor, which can access both dialogue history and response to get ground-truth response emotion. The other is called auxiliary emotion predictor that can only access dialogue history to get anticipated response emotion. Then we use an extra emotion embedding with the predicted results to provide additional inductive bias to GPT2 output. During optimization, we try to reduce the discrepancy between them [11] to fill the gap between training and inference.

Besides, inspired by [12], we treat model training as a deliberation process. Our model is trained for two rounds. The generated responses from the first round could be understood as raw drafts. Then in the second training round, we add the generated responses into dialogue history so that the model can get global information of drafts and polish them up for better generation. To sum up, the contribution in this work are two-fold:

- For EmpDialogue task, we inject anticipated emotion and adopt deliberation training strategy for dialogue generation to obtain better empathetic responses.

* co-corresponding author

- Experimental results[1] illustrate that our method is competitive to other strong baselines on both automatic metrics and human judgment.

2. METHODOLOGY

Given the dialogue history $\{s_0, u_0, s_1, u_1, ..., s_T\}$, where s_* and u_* denote the utterances from speakers and listeners respectively, the goal of EmpDialogue is to generate a empathetic response based on the the dialogue history. In this work, we consider EmpDialogue as a conditional language modeling problem. Figure 1 shows an overview of our proposed model. it mainly consists of two parts: pre-trained GPT2 model [8] as our backbone and two emotion predictors.

EmpDialogue has only provided description of situation with 32 emotion labels for each dialog example, so we have to get emotion labels for each utterance first. According to observation, some emotions have nearly the same meaning with slightly difference. So we cluster these emotions into 7 categories based on their Valence-Arousal-Dominance intensity [13]. And we fine-tune BERT [14] with these emotion training examples to get a strong emotion classifier with 73% accuracy. Then We apply fine-tuned BERT to tag emotion labels for each utterance.

2.1. Input Representation

Essentially, We define the token embedding $T \in \Re^{V \times D}$ and the position embedding P [5]. Meanwhile, to make model could distinguish speaker utterances or listener utterences in multi-turn dialogues, we build dialog state embedding $S \in \Re^{3 \times D}$ and append special end-of-utterance tokens $[eou_1]$/$[eou_2]$ into the speaker/listener utterances for separation. Besides, for better modeling emotion information, we also build the emotion embedding $E \in \Re^{K \times D}$. As shown in Fig 1, we concatenate dialogue history, generated response from first round training and reference response together as whole input for language modeling. For each token w, its final input embedding is:

$$Emb(w) = T(w) + P(w) + E(w) + S(w) \quad (1)$$

2.2. Emotion Injection

Simply applying GPT2 is hard to generate sufficient informative and empathetic responses due to lack of emotion information. To effectively utilize the response emotion, we first propose two kind of response emotions:

- ground-truth emotion (GTEmo): it is the response-aware emotion conditioned on both dialogue history and gold response. Since response itself can be accessed, the predicted result is nearly ground-truth comparing to true emotion label.
- anticipated emotion (ATEmo): it is the response-invisible emotion conditioned only on dialogue history. We want to get the response emotion as accurate as possible with incomplete information.

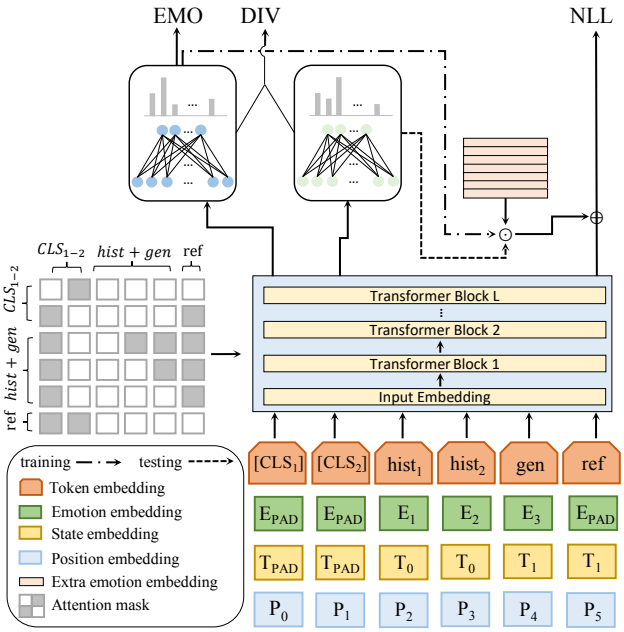

Fig. 1. Model architecture

GTEmo and ATEmo are integrated to GPT2 respectively in training and inference phase.

To achieve this, we add two special tokens $[CLS_1]$ and $[CLS_2]$ at the beginning of all dialogue tokens. As shown in Fig.1, we design a specific attention mask M_{att} to let transformer blocks support different encoding methods. $[CLS_1]$ can only attend to the rightward context and itself so that it gets information from both dialogue history and response. $[CLS_2]$ and dialogue history tokens can attend to each other from both directions so that $[CLS_2]$ can get history contextual representation by bi-directional encoding. Then we obtain hidden output from transformer blocks for both tokens:

$$H_{[cls1]} = TRS(Emb([CLS_1]); M_{att}) \quad (2)$$
$$H_{[cls2]} = TRS(Emb([CLS_2]); M_{att}) \quad (3)$$

After that, $H_{[cls1]}$ is sent to primary emotion predictor to get GTEmo distribution and $H_{[cls2]}$ is sent to auxiliary emotion predictor to get ATEmo distribution. Then we can use the predicted emotion results from $[CLS_1]$ to get an extra emotion embedding for response. However, the argmax operation cannot pass gradient in back-propagation. To deal with non-differentiablity issue, we utilize Gumbel-Softmax trick [15]:

[1]code is available at https://github.com/HLT-HITSZ/EmpGPT

$$logits_{[cls1]} = W_1^T H_{[cls1]} + b_1 \quad (4)$$

$$logits_{[cls2]} = W_2^T H_{[cls2]} + b_2 \quad (5)$$

$$\hat{p}^{(i)} = \frac{\exp\left((logits^{(i)}_{[cls1]} + g^{(i)})\right)/\tau}{\sum_{j=1}^{K} \exp\left((logits^{(j)}_{[cls1]} + g^{(j)})\right)/\tau} \quad (6)$$

where $g^{(i)}$ is from standard Gumbel distribution, i.e: $g^{(i)} = -\log(-\log U^{(i)})$ with $U^{(i)} \sim$ Uniform(0,1), and τ is temperature for better approximate one-hot distribution. K represents emotion category number.

Subsequently, we map the soft predicted emotion category $\hat{p} = [\hat{p}^{(1)}, ..., \hat{p}^{(K)}] \in \Re^{K\times 1}$ into embedded representations by multiplying extra emotion embedding matrix $W_{ext} \in \Re^{K\times D}$. Then we add these embedded representations to the transformer output for gold responses $w_1^{ref}, ..., w_M^{ref}$ (ref part in Fig.1). These hidden outputs were further used for generating the conditional probability of gold responses:

$$H_{w_j^{ref}} = TRS(Emb(w_j^{ref}); M_{att}) + W_{ext}^T \hat{p} \quad (7)$$

$$P(w_j^{ref} | w_{<j}^{ref}, hist, gen) = softmax(W_H^T H_{w_j^{ref}}) \quad (8)$$

During training, we try to minimizing the divergence between GTEmo and ATEmo distribution. This strategy will help auxiliary emotion predictor generating similar feature as primary emotion predictor. As we know, we cannot get access to gold responses in inference phase. So after optimization, we can directly use ATEmo predicted by auxiliary emotion predictor to fill the gap.

2.3. Optimization

Maximum likelihood estimation is embraced widely in language modeling. We adopt NLL loss as follows:

$$\mathcal{L}_{NLL} = -\frac{1}{M}\sum_{i=1}^{M} \log P(w_i^{ref} | w_{<i}^{ref}, hist, gen) \quad (9)$$

As discussed in sec.2.2, emotion classification and distribution divergence optimization are carried out together with language modeling. We first apply softmax operation to Eq.5-6 to get probability distribution $\hat{y}_{[cls1]}$ and $\hat{y}_{[cls2]}$. Then we use cross-entropy loss to optimize emotion classification. Inspired by [11], we utilize soft label $\hat{y}_{[cls1]}$ and hard label together to optimize divergence loss:

$$\mathcal{L}_{EMO} = -y \log \hat{y}_{[cls1]} \quad (10)$$

$$\mathcal{L}_{DIV} = -\beta * y \log \hat{y}_{[cls2]} - \gamma * \frac{\hat{y}_{[cls1]}}{\eta} \log \frac{\hat{y}_{[cls2]}}{\eta} \quad (11)$$

So the final loss function contains three parts. Here α, β and γ are hyper-parameters. η is temperature to smooth distribution:

$$\mathcal{L} = \mathcal{L}_{NLL} + \alpha * \mathcal{L}_{EMO} + \mathcal{L}_{DIV} \quad (12)$$

3. EXPERIMENT

3.1. Experiment Setting

The whole EmpDialogue dataset [3] consists of 25k one-to-one open-domain conversation grounded in different situations which are tagged with 32 emotion labels. We use AdamW [16] optimizer in all of the experiments. For training BERT emotion classifiers, we set batch size with 16 and fine-tune BERT-base model with 5e-5 learning rate for 5 epochs. Then we train our proposed model using 7e-5 learning rate for 8 epochs, and the batch size is set to 32. The value for α, β and γ are 0.1, 0.1, 0.5. We choose the model whose loss function performs best on the validation set as the final model for evaluation. We use topk-topp [17] sampling strategy during decoding where k=20 and p=0.9. Temperature η is set to 5. And τ decreases over iterations via exponential policy: $\tau_n = 1/1000^{n/N}$, where N is maximum training steps and n is current step.

We use several automatic metrics to evaluate the effectiveness of our model, including: average of BLEU-{1,2,3,4} [18] as in [3], embedding score (average, greedy, extrema cosine similarities about word embedding between two sentences) [19], DIST-{1,2} (Count percentage of unique unigrams or bigrams for evaluating diversity of texts) [19], perplexity and average response length. To fully estimate the performance of our model, human evaluations are also conducted. Following [3], we randomly sample 100 dialogues and their corresponding responses. Then 5 human annotators are required to score each aspect (*Empathy*, *Relevance*, *Fluency*). The score range from 1 to 5 (1: not at all, 3: somewhat, 5: very much). A higher score means the better result.

3.2. Baselines

We conduct the experiments with several representative baselines: (1) MoEL: the transformer-based seq2seq model using mixture of several different emotion decoders [4]; (2) EmoPrepend: the transformer-based seq2seq model where the predicted emotion label is prepended to the encoder input [3]; (3) MultiTrans: A simple multi-task transformer trained as [3]; (4) GPT: using GPT2-small to generate required response; (5) MultiGPT: training GPT2-small with simple multi-task strategy similiar to [7, 6]. Ablation study is also implemented for better analyzing our method: (1) ours w/o deli: model with only ATEmo injection; (2) ours w/o emo-inj: model with only deliberate mechanism.

3.3. Results and Analysis

The whole automatic metric results are illustrated in Table 1 and human rating results are listed in Table 2, respectively. Comparing to EmoPrepend, MultiTrans and MoEL, EmoPrepend performs the worst. Although MultiTrans and MoEL get slightly higher BLEU score than ours, the highest DIST

	$\mathrm{BLEU_{avg}}$	DIST-1	DIST-2	$\mathrm{EMB_{gre}}$	$\mathrm{EMB_{ext}}$	$\mathrm{EMB_{avg}}$	PPL	$\mathrm{LEN_{avg}}$
Human	100.00	6.50	38.68	100.00	100.00	100.00	0.00	14.47
EmoPrepend	5.94	0.60	1.98	**70.48**	49.11	81.99	35.11	10.11
MultiTrans	**8.62**	0.41	1.68	69.85	**52.71**	87.93	35.75	11.62
MoEL	8.39	0.61	2.89	69.77	52.39	87.63	37.78	11.46
GPT	5.89	3.18	20.89	68.43	49.91	87.29	15.24	11.87
MultiGPT	6.06	3.20	19.20	68.41	50.25	87.04	13.75	11.17
ours	7.19	3.22	21.00	69.14	50.20	**88.17**	**11.78**	**14.01**
w/o deli	6.71	**3.45**	21.76	68.70	49.86	87.80	13.31	12.71
w/o emo-inj	6.68	3.39	**21.77**	68.49	49.73	87.43	14.09	12.77

Table 1. Automatic metric results on different models. The average, extrema and greedy embedding similarity are abbreviated as $\mathrm{EMB_{avg}}$, $\mathrm{EMB_{ext}}$ and $\mathrm{EMB_{gre}}$. $\mathrm{BLEU_{avg}}$ and $\mathrm{LEN_{avg}}$ are short for average BLEU score and response length.

Model	Relevance	Fluency	Empathy
EmoPrepend	2.56	3.42	2.68
MultiTrans	2.83	3.65	2.89
MoEL	2.74	3.59	2.77
GPT	2.93	3.59	2.93
MultiGPT	3.06	3.66	3.05
ours	**3.50**	**3.96**	**3.52**
w/o deli	3.19	3.57	3.21
w/o emo-inj	3.21	3.68	3.16

Table 2. Results of human ratings on our model and other different baselines

score and lowest perplexity show that our model is superior to others with an extraordinary big step ahead. According to further human evaluation, we find that MoEL and MultiTrans are more likely to generate safe empathetic responses "*I'm sorry*" or "*have a great time*". Average sentence lengths are also shorter than ours. These grammatically correct responses indeed sound reasonable, whereas lack of other useful information or topic-related keywords makes the whole conversation dull and rigid. And to some extent, these kinds of responses cannot cause sufficient emotion resonance and reactions since it is suitable for most situations. For instance, when speakers say "*My daughter is graduating from college. I cannot wait until she finish*", responses of MoEL and MultiTrans are "*I hope you have a great time*" and "*I am sure she will be fine*" respectively. However, our model gets a more informative and relative generation "*Exciting! Do you plan on doing any special events for her?*". Such cases show that our model can effectively elicit empathy and contiguously increase the topic of interests within the communication.

Comparing to GPT-based methods, our model also achieves the highest performance. Longer sentence length and higher BLEU score indicate that our model generates more fluent responses. Our model gets better embedding similarity results and higher DIST scores thanks to directly providing ATEmo information and deliberation process. This also shows that applying the simple multi-task strategy is not powerful enough for modeling emotion. Through ablation study, we find that only applying deliberation process or injecting ATEmo can lead to higher DIST scores but worse embedding similarity results, perplexity and BLEU score. With further human evaluation, we find that two sub-models are likely to generate topic-unrelated or contradictory texts in some situations, which degrade performance in relevance and logicality. For instance, with "*My boyfriend canceled our date last week to go to the movies with his female best friend*" as input, two sub-models generate "*How long did you have the date*" and "*Nice! Are they having fun?*". While our model gets a more empathetic and logical response "*Oh man, that's disappointing, what happened?*".

4. CONCLUSION

In this paper, we propose a novel method that explicitly injecting anticipated emotion into response generation for empathetic dialogue. Through minimizing the divergence between anticipated emotion and ground-truth emotion distribution of responses, the response emotion information are ultilized in both training and inference phase effectively. Besides, deliberation mechanism is also adopted to polish generated results. Experimental results on EmpDialogue dataset show that our model can generate more empathetic and informative responses than other strong baselines.

5. ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (61632011, 61876053, 62006062), the Guangdong Province Covid-19 Pandemic Control Research Funding (2020KZDZX1224), the Shenzhen Foundational Research Funding (JCYJ20180507183527919 and JCYJ20180507183608379), and China Postdoctoral Science Foundation (2020M670912).

6. REFERENCES

[1] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang, "Generating responses with a specific emotion in dialog," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3685–3695.

[2] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," *arXiv preprint arXiv:1704.01074*, 2017.

[3] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5370–5381.

[4] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung, "MoEL: Mixture of empathetic listeners," 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. 121–132.

[5] 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.

[6] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung, "Caire: An end-to-end empathetic chatbot," *CoRR*, vol. abs/1907.12108, 2019.

[7] Rohola Zandie and Mohammad H Mahoor, "Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems," *arXiv preprint arXiv:2003.02958*, 2020.

[8] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," 2019.

[9] J. Shin, P. Xu, A. Madotto, and P. Fung, "Generating empathetic responses by looking ahead the user's sentiment," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7989–7993.

[10] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," in *ACL, system demonstration*, 2020.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[12] Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu, "Deliberation networks: Sequence generation beyond one-pass decoding," in *Advances in Neural Information Processing Systems 30*, pp. 1784–1794. 2017.

[13] Saif Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 174–184.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 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. 4171–4186.

[15] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparametrization with gumble-softmax," in *International Conference on Learning Representations (ICLR 2017)*, 2017.

[16] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2020.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[19] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2122–2132.