Improving Empathetic Response Generation with Retrieval based on Emotion Recognition

Yahui Fu, Koji Inoue, Divesh Lala, Kenta Yamamoto, Chenhui Chu and Tatsuya Kawahara

Abstract Endowing a robot with the ability to express empathy is crucial for building a human-like dialogue system. We propose to improve empathetic response generation with retrieval based on emotion recognition. In addition to a generative model, our model can effectively switch to the retrieval system based on the context and emotion of the user utterances. We incorporate an emotion classifier on top of the context encoder and use context encoding representation to select retrieval responses. Furthermore, it is straightforward to combine our model with the multimodal facial expression of the virtual agent for vivid empathy. Automatic and human evaluations on the Japanese EmpatheticDialogue dataset demonstrate that compared with the solely generative model, our model can generate empathetic responses with more diversity and better scores on the aspects of *Empathy*, *Relevance*, and *Fluency*. Implementing our model on the autonomous android ERICA further demonstrates the effectiveness and adaptivity of our method in achieving an empathetic attentive listening system.

1 Introduction

Incorporating empathy into the dialogue system is essential for improving human-robot interaction experiences, as empathy is the emotional bonding among humans; robots expressing empathy would give humans a feeling of being understood and satisfied with the conversation. Empathetic human-like robots and chatbots have drawn attention in the research community, for example, the human-like android ERICA [1], which has the ability to serve as an empathetic companion during Covid-19 quarantine [2]; Nora, an empathetic dialogue system with a web-based virtual agent in the role of a psychologist [3, 4]; Chatbot CAiRE, which can detect

Graduate School of Informatics, Kyoto University, Japan
e-mail: [fu][inoue][lala][yamamoto][kawahara]@sap.ist.i.kyoto-u.ac.jp, chu@i.kyoto-u.ac.jp

user emotion by textual analysis and respond in an empathetic manner [5]; and Korean mutlimodal empathetic dialogue system, which explored on seven different emotions [6]. However, previous systems express empathy based on coarse-grained emotion recognition or sentiment analysis. In this work, we extend ERICA's ability on attentive listening introduced in our previous research [7] to a fine-grained empathetic system based on elaborate emotion recognition.

Conversational systems based on sequence-to-sequence models [8, 9, 10, 11] encounter the problem of generating safe responses (generic and meaningless, such as "I see") or unnatural responses (have grammatical or logical errors, such as "that is so sweet. I am sorry to hear that"). Previous studies tried to solve this problem by adding external data such as the EmpatheticDialogue dataset or using adversarial learning to mimic humans [12]. However, such approaches are indirect solutions. The retrieval-based models are guaranteed to produce natural and substantial responses, as they are retrieved from external documents, but encounter the problem of producing responses that are not closely relevant to the dialog context. In this study, we incorporate a response retrieval model as a fallback to the generative model based on emotion recognition of 32 categories, which are defined in the EmpatheticDialog dataset. It is difficult to detect emotion from 32 categories accurately, and false emotion detection may mislead the retrieval process. Therefore, we quantify the uncertainty of the emotion predictions as a discriminator to control the response retrieval, which means we only switch to the retrieval when the model is confident about the emotion predicted from the context.

Automatic and human evaluations on the Japanese EmpatheticDialogue dataset have been done to confirm our model can generate empathetic responses with more diversity and better performance on the aspects of *Empathy*, *Relevance*, and *Fluency* compared with the solely generative model. Furthermore, the implementation of our model on the autonomous android robot ERICA demonstrates the effectiveness of our method is achieving an empathetic attentive listening system. In addition, combining our model with the multi-modal facial expression of the virtual agent Gene shows that our model is lightweight and can be straightforwardly implemented for producing vivid empathy.

2 Related Work

2.1 Empathetic Response Generation

Recent conversational systems explore sequence-to-sequence models to generate empathetic responses. For example, Shen et al. [8] proposed a dual-generative model for empathetic response generation based on the assumption that empathy is the emotion consensus between the context and the response. Lin et al. [9] proposed a multi-decoder model that softly combined the outputs of multiple emotion-specific decoders for diverse empathetic response generation. Majumder et al. [10] introduced

an emotion mimicry model on the basis that empathetic responses are often similar to the speaker's emotions. Li et al. [13] proposed an adversarial learning-based model to improve empathy quality, exploiting both dialogue-level and token-level emotions. Li et al. [14] leveraged external commonsense knowledge and emotion lexicon to the Transformer architecture to help emotion detection and empathetic response generation. However, the generative model encounters the problem of generating safe or unnatural responses, and adding external empathetic data does not control quality matched to the context and emotion. To improve empathetic response generation, we incorporate a response retrieval model as a fallback to the generative model based on 32-category emotion recognition, with the emotion uncertainty measure to control the output.

2.2 Retrieval-based Response Generation

Retrieval-based methods have been considered as an alternative or complement to enhance the generation-based approaches. Cai et al. [15] explored a retrieval-guided response generation based on a matching mechanism. Zhang et al. [16] proposed to attentively combine retrieval and generation using a Mixture-of-Experts ensemble to generate a follow-on text. The above studies combined a retrieval system trained with a generation model, thus the effectiveness is very sensitive to the retrieval quality, which may even worsen the generation process. To avoid this problem, we adopt the retrieval system as an alternative to generation based on emotion classification to alleviate the difficulty of empathetic response generation. Specifically, we define 82 empathetic responses for 32 kinds of emotions as a controllable retrieval set, which can be used directly without generation. In addition, when the uncertainty of the emotion predictions is lower than a threshold, our model is confident on the predicted emotion, and then select a response from the retrieval set; otherwise, it outputs the generated response.

2.3 Emotion-driven Response Generation

Integrating emotions into the dialog response generation is significant in building a human-like dialog system. Conventional studies [17, 18] generated empathetic responses conditioning on emotion labels, which is in accordance with the desired emotion but require manually provided emotion labels by the user. Shen et al. [8] incorporated a generation task with emotion classification task to enrich the encoded representations with various emotional traits. However, it did not explicitly feed emotion for generation, as the accuracy of emotion classification is not sufficient, and falsely predicted emotion may misguide the generation process. To avoid this problem and produce an emotion-specific empathetic response, we set the retrieval process

Fig. 1 Snapshot of dialog between ERICA and a subject.

based on emotion recognition and quantify the emotion uncertainty to determine whether to switch to retrieval.

2.4 Attentive Listening System

This work is an extension of the attentive listening system [7] for the autonomous android ERICA [19, 20], which can generate several types of listener responses: backchannels, partial repeats, elaborating questions, assessments, generic sentimental responses, and generic responses. Fig. 1 shows the situation that ERICA is chatting with one subject. Based on the detected sentiment of the user, it can generate positive responses, such as "*That is good (*いいですね*)*" or "*That is nice (*素敵ですね*)*," and negative responses, like "*That is bad (*残念ですね*)*" or "*That is hard (*大変ですね*)*." However, they are stereotypical and not empathetic enough. To improve the human-robot interaction experiences, we extend previous sentiment responses based on positive/negative sentiments to empathetic responses based on 32 emotion categories.

Furthermore, in daily conversation, neutral utterances account for the largest percentage than empathetic expressions; it would be weird if the system always outputs empathetic responses. Therefore, we only switch to retrieve an empathetic response when emotion uncertainty is low; for example, if the emotion of input is neutral, which is not included in our 32 emotions, then the emotion uncertainty would

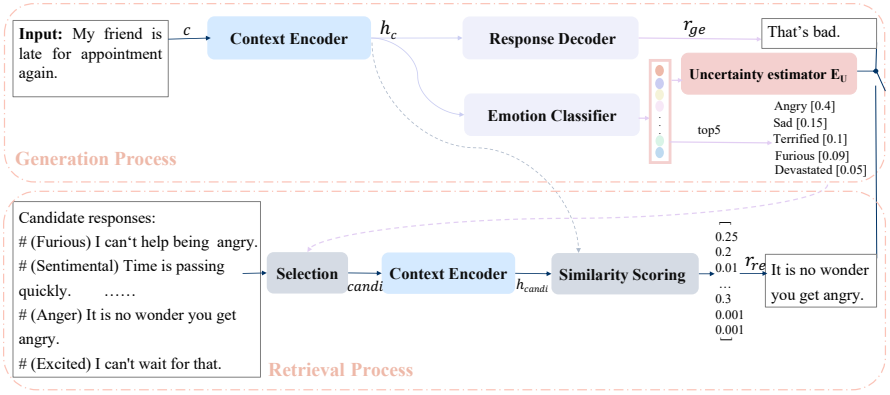


Fig. 2 The proposed model which calibrates generation by an alternative retrieval system based on emotion recognition, where generation and retrieval processes share the context encoder.

be very high; in this case, we will switch to the other kinds of neutral responses like partial repeats as listed in the previous attentive listening system.

3 Proposed Model

As shown in the Fig. 2, for a given context c, the goal is to generate an empathetic response r_{ge} by a sequence-to-sequence generative model and retrieve a response r_{re} based on the predicted emotion. We also quantify the uncertainty of the predicted emotion E_U; if it is smaller than the predetermined threshold, then take the retrieved response as the final output; otherwise, use the generated response.

3.1 Context Encoder and Response Decoder

Inspired by [8, 12], we employ the dual-learning mechanism and utilize Transformer as the sequence-to-sequence base model for the generation process. We add a special token CLS and SOS to the beginning of the encoder input and decoder input respectively, which represents the global memory of the whole sequence.

$$\begin{aligned} h_c &= context_{enc}(emb_w(c)) \\ h_r^t &= response_{dec}[emb_w(r_{ge}^{<t}), h_c, z] \end{aligned} \tag{1}$$

where emb_w represents word embedding layer, $context_{enc}$, $response_{dec}$ are the Transformer-based encoder and decoder, respectively. $h_c \in \mathbb{R}^{n \times d_{enc}}$ and $h_r^t \in \mathbb{R}^{n \times d_{dec}}$ are encoded context representation and decoded response representation, respec-

tively; n is the number of encoder/decoder layer, and d_{enc}, d_{dec} is the dimension of the encoder and decoder layer, respectively; t means t-th token among the generated response. z is the Gaussian distribution [21], to capture consensus between the context and response.

$$\begin{aligned} z &= \mu_c + \sigma_c \odot \varepsilon \\ \varepsilon &\sim \mathcal{N}(0, I) \end{aligned} \tag{2}$$

And μ and σ are computed based on the context encoder outputs h_c:

$$\begin{aligned} \mu_c &= \omega_1 h_c + b_1 \\ \sigma_c^2 &= \omega_2 h_c + b_2 \end{aligned} \tag{3}$$

where $\omega_1, \omega_2, b_1, b_2$ represent feedforward network weights and biases. Then, the generation distribution over the vocabulary of the next token is:

$$p(y_t) = softmax(W_v h_r^t + b_v) \tag{4}$$

W_v, b_v are the weights and a bias of the corresponding softmax network.

3.2 Emotion Classifier

We introduce an emotion classifier on top of the context encoder to explicitly detect the emotion from the user utterance. We compute the emotion distributions as below:

$$p_e = softmax(W_e h_{c_0} + b_e) \tag{5}$$

where W_e, b_e are the weights and a bias of the corresponding emotion classifier network. We take the CLS embedding h_{c_0} of the encoder output as the representations of the entire context.

To train the sequence-to-sequence model and emotion classifier, given the paired data (c, y) and the speaker emotion state e_s, we minimize the negative log-likelihood of generating the ground truth target sequence $y = (y_1, y_2, ..., y_M)$, as well as the emotion probability with the cross-entropy loss function:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}e_s + \mathcal{L}gen \\ \mathcal{L}e_s &= -\log(p_{e_s}) \\ \mathcal{L}gen &= \sum_{t=1}^{M} -\log p(y_t | y_{<t}, c) \end{aligned} \tag{6}$$

where M is the length of a ground truth response.

3.3 Retrieval Process

To alleviate the difficulty of generating suitable empathetic responses, we incorporate the retrieval process to serve as an alternative of the generation process as shown in Fig. 2. We first compute the emotion distributions of the input context as shown in Equation (5). Then, we select the corresponding n candidate responses from our pre-defined data set based on the predicted emotions. We use the same context encoder to encode the selected candidate responses:

$$h_{candi_i} = trs_{enc}(candi_i) \tag{7}$$

where $candi_i$ is the i-th selected candidate response, and i ranges from 1 to n. Then, we compute the similarity score $sim_{i,c}$ between the candidate representation h_{candi_i} and input context h_c:

$$sim_{i,c} = 1 - \arccos\left(\frac{h_c^\top h_{candi_i}}{\|h_c\| \left\|h_{candi_i}\right\|}\right)/\pi. \tag{8}$$

Then, candidate r_{re} is chosen by the ranking of similarity score.

3.4 Uncertainty Estimator

We quantify the uncertainty of the emotion prediction, computed by the entropy of the emotion classification probabilities:

$$E_U = \sum_{v=1}^{V} p_e^v \log p_e^v \tag{9}$$

where V is the number of the emotion categories. After obtaining the generated response r_{ge} and retrieved response r_{re}, we choose the suitable one based on a threshold u:

$$r = \begin{cases} r_{re}, & \text{if } E_U < u \\ r_{ge}, & \text{if } E_U \geq u \end{cases} \tag{10}$$

4 Experimental Evaluations

4.1 Dataset

We train our model based on the Japanese version EmpatheticDialogues [22], which is created by following the original EmpatheticDialogues database [23]. Japanese

native speakers are engaged for constructing situation sentences and dialogues. Each dialog contains four utterances by two persons interacted in the form as "ABAB." We train and evaluate our model for each turn of *Listener (B)* responding to *Speaker (A)*, and extend *Speaker (A)*'s inquiries one by one from context histories. There are 20,000 dialogues in total with 32 evenly distributed emotion labels, and utterances in each dialogue share the same emotion label. The ratio for training/validation/test set is 8:1:1. In addition, we further test our model on a human-robot dialogue corpus where an android ERICA talked with 20 elderly people in the task of attentive listening [7].

For the retrieval process, a Japanese speaker created two or three candidate responses for each emotion category that can be used in many situations and do not depend on the context. In total, there are 82 candidate responses.

4.2 Settings

We set the batch size to 16 and the learning rate to 0.0001. We used JUMAN++ for Japanese word segmentation. We used pre-trained fastText [24] vectors to initialize the word embeddings. All hyper-parameters of the Transformer model were set according to the previous work [11]. We used greedy search during inference in the generation process and the maximum decoding step was set to 30.

4.3 Comparison Models

For a comprehensive evaluation, we compare our model with other state-of-the-art models.

Transformer [23]: This is a standard transformer encoder-decoder architecture model. After encoder, it coupled a response decoder and emotion classification.

MoEL [9]: An extension of Transformer, which softly combines multiple emotion-specific decoders to a meta decoder to generate an empathetic response.

MIME [10]: This method assumed that empathetic responses often mimic the speaker's emotion and integrated emotion grouping, emotion mimicry, and stochasticity into the emotion mixture for various empathetic responses.

4.4 Evaluation Measures

4.4.1 Automatic Metrics

For automatic evaluation, we use the following metrics: (1) BLEU [25] which evaluates the matching of the generated response to the ground truth. We use *multi-*

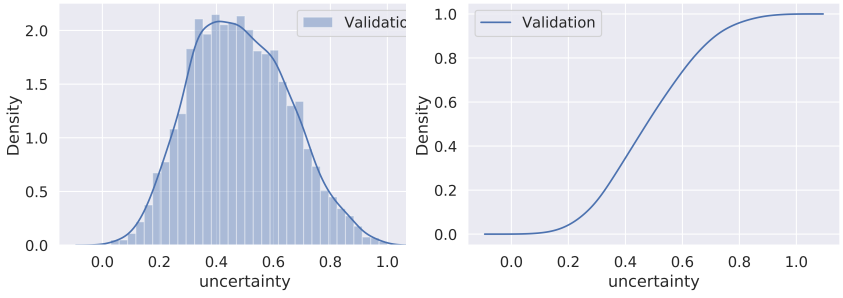
bleu.perl [26] to compute the BLEU scores. (2) Dist-1/Dist-2 (Distinct-1/ Distinct-2) [27] to evaluate the diversity of the generated response.

4.4.2 Human Evaluation

We randomly sample 100 dialogues and their corresponding responses generated from our method as well as the compared methods. We recruit crowd-workers to evaluate the responses generated by various models. Annotators are asked to evaluate the quality of the generated response based on three dimensions: Empathy, Relevance, and Fluency [8, 10, 23]. Three crowd-workers evaluate each dimension, and we take the average value. Empathy measures whether the generated response contains the emotion understanding of the context. Relevance considers the topic consistency between the context and generated response. Fluency assesses whether the generated responses are grammatically correct and readable. Each metric is rated on a scale from 1 to 5.

4.5 Emotion Uncertainty Threshold

It is important to find a suitable threshold for the emotion uncertainty estimator to select the final output from the generated and retrieved responses. Figure 3 depicts the count and cumulative distributions of the emotion uncertainty in the validation set. For example, we can see from the cumulative distribution that there is about 18% percent of the samples with emotion uncertainty smaller than 0.3, which means if the emotion uncertainty threshold is set to 0.3, 18% percent of generated responses will be replaced by the corresponding retrieved one. Based on the values of *D1* and *D2* in Table 1, we choose the emotion uncertainty threshold to be 0.3 or 0.4.



(a) Count distribution (b) Cumulative distribution

Fig. 3 Emotion uncertainty distribution on the validation set

Table 1 Results of the proposed method with different uncertainty values on the validation set of Japanese EmpatheticDialogues dataset.

Uncertainty	Cumulative	Dist-1(%)	Dist-2(%)
0.2	0.05	2.08	8.01
0.3	0.18	2.21	**8.31**
0.4	0.35	**2.29**	8.25
0.5	0.50	2.28	8.06

5 Results and Analysis

5.1 Results on Japanese EmpatheticDialogue dataset

The effectiveness of the alternative retrieval process is shown in Table 2 using the test set. Compared with the comparative baselines and our generative model, both *Generation + Retrieval (E_u=0.3)* and *Generation + Retrieval (E_u=0.4)* are superior in the automatic evaluation aspects of *Dist-1*, *Dist-2*, and human evaluation dimensions of *Empathy*, *Relevance*, and *Fluency*, but inferior in the aspect of *BLEU*. However, it has been argued that *BLEU* is inappropriate to measure the quality of empathetic responses, as response generation is not the task with a unique solution [12, 15]. The result demonstrates the effectiveness of our model to generate empathetic responses with better diversity and higher score by human evaluation. It confirms the validity that applies the plug-and-play retrieval process as an alternative to the generation of the method based on the emotion uncertainty estimation.

In addition, we can see that the emotion uncertainty threshold set to 0.3 is superior to 0.4 in the aspects of *Empathy* and *Relevance*, inferior in *Fluency*. We can learn from it in two points; firstly, emotion uncertainty set to 0.3 is optimal for our model to integrate generation with retrieval process when experimented on the Japanese EmpatheticDialogue dataset; secondly, the retrieval process has a significant advantage in the part of *Fluency* because the retrieval set is pre-created by a native speaker in advance, which has the potential to solve the problem of generating unnatural responses in a generative model.

Table 2 Evaluation of the proposed and baseline methods for the test set of Japanese EmpatheticDialgues dataset.

Model	Automatic Evaluation			Human Evaluation		
	BLEU	Dist-1 (%)	Dist-2 (%)	Empathy	Relevance	Fluency
Transformer [23]	**6.92**	1.34	5.77	2.88	2.47	2.89
MoEL [9]	0.66	1.36	5.67	3.15	2.74	2.95
MIME [10]	0.64	0.69	2.62	3.22	2.77	3.24
Generation	6.79	2.06	7.94	3.47	3.22	3.24
Generation + Retrieval (E_u=0.3)	5.74	2.18	**8.14**	**3.67**	**3.41**	3.71
Generation + Retrieval (E_u=0.4)	6.23	**2.20**	7.99	3.50	3.29	**3.96**

Table 3 Case studies of empathetic attentive listening system

case	*Input*	Baseline	Retrieval system	Predicted emotion
1	後々不安ということでね。 I mean, I'm worried about it later on.	大変ですね。 It's hard.	不安ですね。 I'm worried about it.	不安に思う anxious
2	怖いからなんかこうした時に玄関の靴を 持って逃げ出したりしたことがあったんです。 I used to run away with the shoes on the front porch when I was scared of something like this.	大変ですね。 It's hard.	恐怖を感じますね。 I feel scared.	怖い terrified
3	あまりお付き合いできないのは, もう大変残念で悲しいこと です。 It's a great pity and a great sadness that we can't socialise more.	大変ですね。 It's hard.	悲しいですね。 That's sad.	悲しい sad
4	一緒に話したりするのとはちょっと感覚が 違ったから、多少孤独を感じた場面あったかな。 It felt a bit different from talking together, so there were times when I felt a bit lonely.	残念でしたね。 That was bad.	センチメンタルですね。 That's sentimental.	感傷的になる sentimental
5	やっぱ どうしても、みんなから ちょっと嫌な目で見られたりと か。 I can't help but get a bit of a disapproving look from people.	残念でしたね。 That was bad.	それは嫌な気持ちに なりますね。 That's a disgusting feeling.	嫌悪感を抱く disgusted
6	雑談というか言うて元気な姿を見せる のが一番。 The best way is to chat with them, or rather, say something and show them how healthy you are.	素敵ですね。 It's nice.	気分いいですね。 Feels great.	楽しい joyful
7	それは感激する旅行でございました。 It was an inspiring trip.	素敵ですね。 it's nice.	充実してますね。 It's very fulfilling.	満足 content
8	外国旅行いろいろ楽しい思い出は いろいろ思い出してきているんで I have many pleasant memories of traveling abroad.	いいですね。 It's good.	楽しそうですね。 Sounds like a lot of fun.	楽しい joyful
9	ほんとに毎週土曜日が楽しみだったんです。 I really looked forward to it every Saturday.	いいですね。 It's good.	待ち遠しいですね。 I can't wait for it.	わくわくする excited
10	あんまり楽しみくなりましてね。 I'm really looking forward to it.	いいですね。 it's good.	楽しみですね。 I'm looking forward to it.	期待する anticipating

5.2 Case Studies of Empathetic Attentive Listening System

We further implement our retrieval model to the previous attentive listening system [7] for the autonomous android ERICA [1] to make an empathetic attentive listening system. For each input utterance, we recognize its emotion and compute emotion uncertainty; if it is lower than the pre-defined threshold, we select one response by the retrieval system; otherwise use the responses like backchannels, partial repeats as defined in [7] to continue the conversation.

We present a comparation between our retrieval system and sentimental responses in the previous attentive listening system in Table 3. We can see that compared with the previous system, which produces four kinds of sentimental responses based

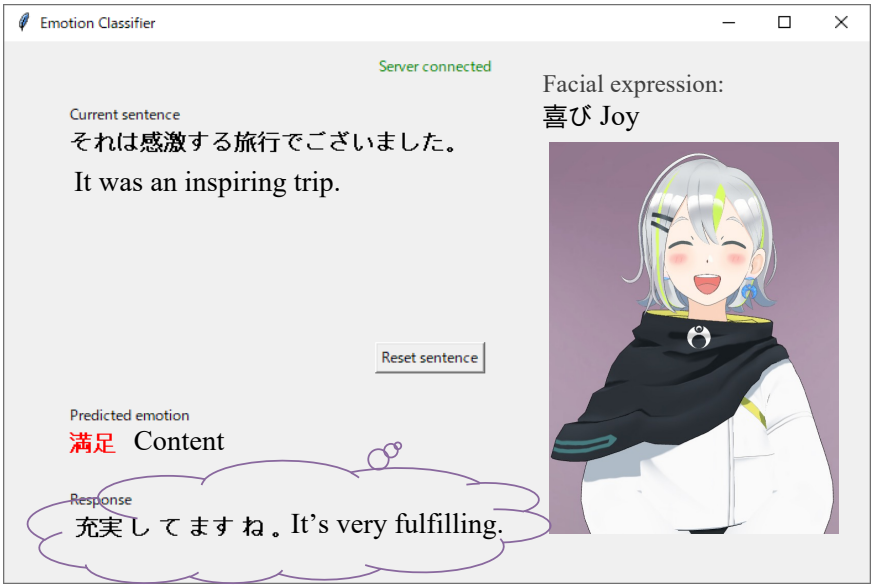


Fig. 4 Example of response with multi-modal facial expression of the virtual agent Gene.

on positive and negative sentiments, our retrieval system can generate fine-grained empathetic responses corresponding to different detected emotions. Therefore, our model can improve human-robot interaction experiences.

5.3 Combination with Multi-modal Facial Expression

In a human-robot interaction system for a multi-modal expressions, such as audio, facial and motion, can significantly improve user experiences. We further combine our model with the facial expression for a virtual agent Gene [28] to produce vivid empathy. We can see from Fig. 4 that for the input utterance "それは感激する旅行でございました。 (It was an inspiring trip)", the system has an emotion prediction as "満足 (content)" and its emotion uncertainty is lower than the selection threshold, so the system outputs retrieved response as "充実してますね。 (It's very fulfilling.)", as well as a "喜び(joy)" facial expression.

6 Conclusion

In this paper, we have proposed calibrating a generative model with a retrieval system based on emotion recognition to improve empathetic response generation.

We replace generation with retrieval in an optimal ratio to combine the advantages of generative and retrieval models, which was determined according to the response diversity of the system on the validation set. Automatic and human evaluations on the Japanese EmpatheticDialogue dataset illustrate that compared with the solely generative model, our model can generate empathetic responses with more diversity and better scores on the aspects of *Empathy*, *Relevance*, and *Fluency*. Implementing our model on the autonomous android ERICA further demonstrates the effectiveness and adaptivity of our method in achieving an empathetic attentive listening system. In addition, combining our model with the multi-modal facial expression of the virtual agent Gene shows that our model is lightweight, which can be straightforwardly implemented for producing vivid empathy. We will explore emotion causes in the dialog for reasonable empathetic response generation in our future work.

Acknowledgements This work was supported by JSPS KAKENHI (JP19H05691) and JST Moonshot R&D (JPMJPS2011). This work was also supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2123. The agent system described in Section 5.3 was provided by Akinobu Lee (Nagoya Institute of Technology).

References

[1] Glas, D., Minato, T., Ishi, C., Kawahara, T. & Ishiguro, H. Erica: The ERATO intelligent conversational android. *2016 25th IEEE International Symposium On Robot And Human Interactive Communication (RO-MAN)*. pp. 22-29 (2016)

[2] Ishii, E., Winata, G., Cahyawijaya, S., Lala, D., Kawahara, T. & Fung, P. Erica: an empathetic android companion for COVID-19 quarantine. *ArXiv Preprint ArXiv:2106.02325*. (2021)

[3] Winata, G., Lovenia, H., Ishii, E., Siddique, F., Yang, Y. & Fung, P. Nora: The well-being coach. *ArXiv Preprint ArXiv:2106.00410*. (2021)

[4] Winata, G., Kampman, O., Yang, Y., Dey, A. & Fung, P. Nora the Empathetic Psychologist.. *INTERSPEECH*. pp. 3437-3438 (2017)

[5] Lin, Z., Xu, P., Winata, G., Siddique, F., Liu, Z., Shin, J. & Fung, P. Caire: An end-to-end empathetic chatbot. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **34**, 13622-13623 (2020)

[6] Jung, M., Lim, Y., Kim, S., Jang, J., Shin, S. & Lee, K. An Emotion-based Korean Multimodal Empathetic Dialogue System. *Proceedings Of The Second Workshop On When Creative AI Meets Conversational AI*. pp. 16-22 (2022)

[7] Inoue, K., Lala, D., Yamamoto, K., Nakamura, S., Takanashi, K. & Kawahara, T. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. *Proceedings Of The 21th Annual Meeting Of The Special Interest Group On Discourse And Dialogue*. pp. 118-127 (2020)

[8] Shen, L., Zhang, J., Ou, J., Zhao, X. & Zhou, J. Constructing Emotional Consensus and Utilizing Unpaired Data for Empathetic Dialogue Generation.

Findings Of The Association For Computational Linguistics: EMNLP 2021. pp. 3124-3134 (2021)

[9] Lin, Z., Madotto, A., Shin, J., Xu, P. & Fung, P. MoEL: Mixture of Empathetic Listeners. *EMNLP-IJCNLP*. pp. 121-132 (2019)

[10] Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A., Mihalcea, R. & Poria, S. MIME: MIMicking Emotions for Empathetic Response Generation. *EMNLP*. pp. 8968-8979 (2020)

[11] Sabour, S., Zheng, C. & Huang, M. CEM: Commonsense-aware Empathetic Response Generation. *ArXiv Preprint ArXiv:2109.05739*. (2021)

[12] Cui, S., Lian, R., Jiang, D., Song, Y., Bao, S. & Jiang, Y. Dal: Dual adversarial learning for dialogue generation. *ArXiv Preprint ArXiv:1906.09556*. (2019)

[13] Li, Q., Chen, H., Ren, Z., Ren, P., Tu, Z. & Chen, Z. EmpDG: Multiresolution interactive empathetic dialogue generation. *ArXiv Preprint ArXiv:1911.08698*. (2019)

[14] Li, Q., Li, P., Ren, Z., Ren, P. & Chen, Z. Knowledge Bridging for Empathetic Dialogue Generation. *ArXiv E-prints*. pp. arXiv-2009 (2020)

[15] Cai, D., Wang, Y., Bi, W., Tu, Z., Liu, X. & Shi, S. Retrieval-guided dialogue response generation via a matching-to-generation framework. *Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing (EMNLP-IJCNLP)*. pp. 1866-1875 (2019)

[16] Zhang, Y., Sun, S., Gao, X., Fang, Y., Brockett, C., Galley, M., Gao, J. & Dolan, B. RetGen: A Joint framework for Retrieval and Grounded Text Generation Modeling. (2022)

[17] Zhou, H., Huang, M., Zhang, T., Zhu, X. & Liu, B. Emotional chatting machine: Emotional conversation generation with internal and external memory. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **32** (2018)

[18] Colombo, P., Witon, W., Modi, A., Kennedy, J. & Kapadia, M. Affect-driven dialog generation. *ArXiv Preprint ArXiv:1904.02793*. (2019)

[19] Inoue, K., Milhorat, P., Lala, D., Zhao, T. & Kawahara, T. Talking with ERICA, an autonomous android. *Proceedings Of The 17th Annual Meeting Of The Special Interest Group On Discourse And Dialogue*. pp. 212-215 (2016)

[20] Kawahara, T. Spoken dialogue system for a human-like conversational robot ERICA. *9th International Workshop On Spoken Dialogue System Technology*. pp. 65-75 (2019)

[21] Kingma, D. & Welling, M. Auto-Encoding Variational Bayes. *ArXiv Preprint ArXiv:1312.6114*. (2013)

[22] Sugiyama, H., Mizukami, M., Arimoto, T., Narimatsu, H., Chiba, Y., Nakajima, H. & Meguro, T. Empirical Analysis of Training Strategies of Transformer-based Japanese Chit-chat Systems. *ArXiv Preprint ArXiv:2109.05217*. (2021)

[23] Rashkin, H., Smith, E., Li, M. & Boureau, Y. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. *ACL*. pp. 5370-5381 (2019)

[24] Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *TACL*. **5** pp. 135-146 (2017)

[25] Papineni, K., Roukos, S., Ward, T. & Zhu, W. Bleu: a method for automatic evaluation of machine translation. *ACL*. pp. 311-318 (2002)
[26] Britz, D., Goldie, A., Luong, M. & Le, Q. Massive Exploration of Neural Machine Translation Architectures. *EMNLP*. pp. 1442-1451 (2017)
[27] Li, J., Galley, M., Brockett, C., Gao, J. & Dolan, B. A Diversity-Promoting Objective Function for Neural Conversation Models. *NAACL-HLT*. pp. 110-119 (2016)
[28] Lee, A. & Ishiguro, H. Development of CG-based Embodied Dialogue Agents and System with Conversational Reality for Avatar-Symbiotic Research. *SIG-SLUD, JSAI*. (2022)