Dual Variational Generative Model and Auxiliary Retrieval for Empathetic Response Generation by Conversational Robot

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

Graduate School of Informatics, Kyoto University, Japan

ARTICLE HISTORY
Compiled February 28, 2023

ABSTRACT

Empathy in human-robot conversations aims to endow the robot with the ability to comprehend user emotion and experience, and then respond to it appropriately. Generally, empathy is embodied in the aspects of both contextual understanding and affective expression, which occur when there exist content and emotion consistencies between context and response. However, previous studies only focus on either aspect. In this paper, we propose a dual variational generative model (DVG) for empathetic response generation to achieve both. Specifically, we integrate an emotion classifier and a variational autoencoder (VAE) into a dual response and context generative model to learn the emotion and content consistencies efficiently. DVG utilizes VAE to mimic the process of context/response understanding. In addition to the generative model, our model can effectively switch to another retrieval system as a fallback solution. Automatic and human evaluations on Japanese and English EmpatheticDialogue datasets demonstrate the effectiveness of our method for empathetic response generation. Furthermore, we evaluate our model's ability in general response generation, which is not specific to empathetic but also chitchatting dialogue system.

KEYWORDS
Conversational robot; Spoken dialogue system; Empathetic response generation; VAE

1. Introduction

Empathy is a desirable capacity of humans to place themselves in another's position to show understanding of his/her experience and feelings and respond appropriately. It has been widely argued that empathetic responding contributes to better human-machine interaction experience and satisfaction in a wide range of domains, such as medical therapeutics [1,2] and social chatbots [3,4]. Therefore, generating empathetic dialogue responses is of great significance for conversational robots.

In general, empathy includes aspects of contextual understanding and affection [5], which represent perceiving the user's situation and expressing emotion, such as the 'Empathetic response' shown in Figure 1. However, previous studies either focused on detecting user emotion and embedding emotional traits to generate responses with affection [6–8], or focused on integrating commonsense knowledge to help contextual

CONTACT Yahui Fu Email: fu.yahui.64p@st.kyoto-u.ac.jp

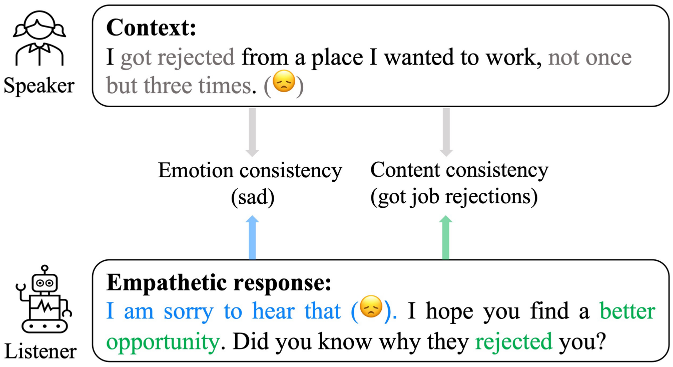

Speaker
Context:
I got rejected from a place I wanted to work, not once but three times. (😞)
Emotion consistency (sad)
Content consistency (got job rejections)
Listener
Empathetic response:
I am sorry to hear that (😞). I hope you find a better opportunity. Did you know why they rejected you?

Figure 1. An example of an empathetic response from the JEmpatheticDialogue dataset [10]. Blue highlighted text denotes the affective expression, and green text implicates context understanding.

understanding [9]. To make a further exploration on both aspects for empathetic response generation, in this paper, we propose a dual variational generative (DVG) model.

Our DVG model is based on the assumption that there exist emotion and content consistencies between context and appropriate empathetic response, as shown in Figure 1. To capture such consistency, we utilize the mutual information from the duality of response generation and context generation. Specifically, we introduce a variational autoencoder (VAE) into the dual generative model to mimic the process of context/response understanding by reconstruction and utilize an emotion classifier to capture the emotion state during the conversation for affective expression. These will enhance the shared variational variables of the dual generative model with content and emotion consistencies.

The generative models can produce an empathetic response, but, they 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'). Instead, the retrieval-based models are guaranteed to produce natural and empathetic responses, as they are retrieved from external documents, but encounter the problem of producing responses that are not closely relevant to the dialog context. Therefore, we incorporate a response retrieval model as a fallback to the generative model based on emotion recognition to leverage the merits of both the generative and retrieval model. Specifically, we define 82 empathetic responses conditioned on 32 kinds of emotions as a controllable retrieval set. It is difficult to detect emotion 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.

In daily-style conversation, an empathetic response is just one kind of conversation reply, while neutral chatting also accounts for a large percentage. Therefore, we further enrich our model's ability to build a general Japanese dialogue system by incorporating the daily life dataset *PersonaChat* [10] into training.

Our main contributions are summarized as follows:

- We propose a DVG model to efficiently learn the bidirectional relationship between the context and the response in the conversation for contextual understanding and affective expression. Automatic and human evaluations on both

Japanese and English EmpatheticDialogue datasets show that our method outperforms competitive baselines.
- We introduce a retrieval system as a fallback to the generation process to directly produce an empathetic response. Automatic and human evaluations on the Japanese EmpatheticDialogue dataset demonstrate that compared with the solely generative model, our generative+retrieval system can generate empathetic responses with more diversity and better scores on the aspects of *Empathy*, *Relevance*, and *Fluency*.
- We evaluate our method in general response generation, which is not specific to empathetic but also chitchatting dialogue system. Automatic evaluation and human-agent interaction experiments further demonstrate our system's effectiveness.

2. Related Work

2.1. Empathetic Response Generation

Existing studies for empathetic response generation are mostly based on either affection or contextual understanding. Lin et al. [6] softly combined the output of multiple emotion-specific decoders to improve appropriate empathetic response generation. Majumder et al. [7] argued that empathetic responses often mimic the speaker's emotion, then proposed emotion grouping and emotion mimicry to generate empathetic and various responses. Sabour et al. [9] leveraged ATOMIC [11], which is a knowledge base of commonsense reasoning inferences about if-then events to improve contextual understanding in the dialog.

Different from the previous studies, we not only focus on affective expression using a dual generative model with emotion classifier, but also utilize VAE to force the content consistency for the aim of contextual understanding.

2.2. Dual Learning

Dual learning has been applied to several tasks due to its potential in improving the performance of both the primary task and auxiliary task. Tseng et al. [12] coupled natural language understanding and natural language generation through a shared latent variable, which benefits both tasks. Cui et al. [13] utilized the additional information from a response to query generation to avoid safe response. Hu et al. [14] integrated bidirectional learning with a discriminator for neural topic modeling.

In this study, we extend dual learning to efficiently learn the bidirectional relationship between context and response.

2.3. 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

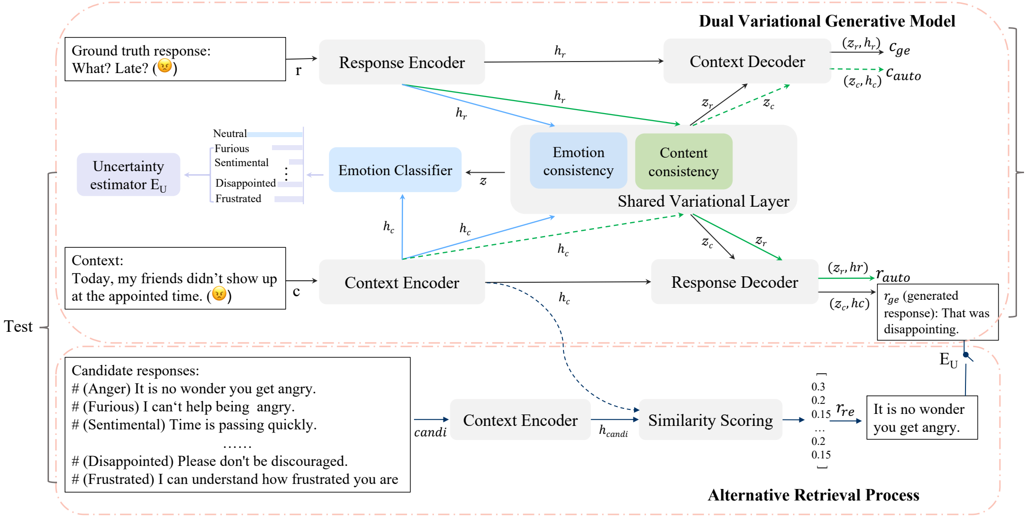

Dual Variational Generative Model
Ground truth response:
What? Late? (😠)
r
Response Encoder
h_r
Context Decoder
(z_r, h_r) c_{ge}
(z_c, h_c) c_{auto}
z_r
z_c
Neutral
Furious
Sentimental
Disappointed
Frustrated
Uncertainty estimator E_U
Emotion Classifier
z
Emotion consistency
Content consistency
Shared Variational Layer
Train
Test
h_c
Context:
Today, my friends didn't show up at the appointed time. (😠)
c
Context Encoder
Response Decoder
(z_r, hr) r_{auto}
(z_c, hc)
r_{ge} (generated response): That was disappointing.
E_U
Candidate responses:
# (Anger) It is no wonder you get angry.
# (Furious) I can't help being angry.
# (Sentimental) Time is passing quickly.
……
# (Disappointed) Please don't be discouraged.
# (Frustrated) I can understand how frustrated you are
candi
Context Encoder
h_{candi}
Similarity Scoring
0.3
0.2
0.15
…
0.2
0.15
r_{re}
It is no wonder you get angry.
Alternative Retrieval Process

Figure 2. Proposed DVG model for empathetic response generation. Blue and green highlighted lines and blocks denote the emotion consensus and content consistency processes separately. Specifically, green solid lines represent the response reconstruction process, while green dotted lines represent context reconstruction. c_{ge}, c_{auto}, r_{auto}, r_{ge} mean context generation, context auto-reconstruction, response auto-reconstruction, and response generation, respectively. Compared with previous studies, we incorporate variational decoder to the dual generative model for context and response reconstruction (c_{auto} and r_{auto}).

retrieval system as a fallback to the generation model based on emotion classification to alleviate the difficulty of empathetic response generation.

3. Baseline Dual Generative Model

Our proposed DVG model is based on a dual generative model, which coupled the response generation from context and context generation from response with one duality layer. The duality layer models the mutual relationships between the context and response, such as emotion consensus [8]. The basic unit of generation module can be chosen from GRU [13], LSTM [12] and Transformer [8]. In this work, we utilize Transformer [17] encoder and decoder.

Shen et al. [8] tried to ensure the emotion consistency from duality complementarity (with the blue circle in Figure 2) and composed the shared layer with a simple dense and softmax networks. However, the variables of the dense layer are deterministic. In this paper, we design the shared layer to be variational, which allows for composing random variables to generate diverse responses. We also incorporate VAE into the dual generative model with a reconstruction process (e.g. context to context) to enhance the shared layer for better content consistency, in addition to the consistency between the context and response.

4. Dual Variational Generative (DVG) Model

As shown in the green solid and dotted flows in Figure 2, we incorporate a VAE into the dual generation framework. The variational decoder is utilized for not only generation but also reconstruction between the context and response, and the reconstruction process ensures the consistency and makes the learning of the shared layer easy. For a given context c, the goal is to generate an empathetic response r_{ge} by the proposed DVG. We will explain each module in the following.

4.1. Model Architecture

There are two similar processes in our DVG model. One is the forward dialogue process from context to response (r_{ge}): context encoder, shared variational layer, and response decoder. The other is the backward dialogue process from response to context (c_{ge}): response encoder, shared variational layer, and context decoder. Moreover, we incorporate the variational auto-reconstruction process from context to context (c_{auto}) and from response to response (r_{auto}). We utilize Transformer for the encoders and decoders. An emotion classifier is augmented to this model. We describe the details of the forward dialogue process in this subsection.

4.1.1. Context Encoder

Inspired by Devlin et al. [18], we firstly add a special token CLS to the beginning of the context c, which represents the global memory of the whole sequence. Then the input context c are converted to word embeddings $emb_w(c)$, summed with the position embeddings $emb_{pos}(c)$:

$$e_c = emb_w(c) + emb_{pos}(c) \tag{1}$$

Finally, we employ Transformer encoder to get the context representation:

$$h_c = trs_{enc_c}(e_c) \tag{2}$$

where trs_{enc_c} is the forward context encoder, $h \in \mathbb{R}^{n \times d_{enc_c}}$, n is the number of encoder layers, and d_{enc_c} is the dimension of the encoder layer.

4.1.2. Shared Variational Layer

We assume that there exists a continuous latent representation z, which represents the mutual characteristics, underlying a pair of context c and response r, where z can be inferred from either c or r. Considering the intractable posterior distribution of unobserved variable z, inspired by Kingma et al. [19], we choose the posterior distribution $q_{trs_{enc_c}}(z|h_c)$ to be Gaussian, trs_{enc_c} is the forward context encoder, and utilizes the reparameterization trick:

$$\begin{aligned} z_c &= \mu_c + \sigma_c \odot \epsilon \\ \epsilon &\sim \mathcal{N}(0, I) \end{aligned} \tag{3}$$

Then we use the hidden layer of the context encoder output h_c to compute the variable μ and σ in the variational process:

$$\begin{aligned} \mu_c &= \omega_1 h_c + b_1 \\ \sigma_c^2 &= \omega_2 h_c + b_2 \end{aligned} \tag{4}$$

where $\omega_1, \omega_2, b_1, b_2$ represent feedforward network weights and biases. For the backward response model, we do the same process:

$$\begin{aligned} z_r &= \mu_r + \sigma_r \odot \epsilon \\ \mu_r &= \omega_3 h_r + b_3 \\ \sigma_r^2 &= \omega_4 h_r + b_4 \end{aligned} \tag{5}$$

where h_r is the output of response encoder trs_{enc_r} in the backward response model, $\omega_3, \omega_4, b_3, b_4$ represent backward network weights and biases.

4.1.3. Response Decoder

We incorporate the mutual representation z into the Transformer decoder [17] for output generation. First, we add a special token SOS to the beginning of the decoder input $y_{<t}^{(i)}$, and conduct word and positional embeddings:

$$e_t^{(i)} = emb_w(y_{<t}^{(i)}) + emb_{pos}(y_{<t}^{(i)}) \tag{6}$$

where i is in a value index of $\{r, c\}$. To efficiently learn the representation of z_c, used for response generation, we also introduce a task of context reconstruction, which involves contextual understanding, inspired by VAE [19].

$$\begin{aligned} r_{ge} &= trs_{dec_r}([e_t^{(r)}, h_c, z_c]) \\ c_{auto} &= trs_{dec_c}([e_t^{(c)}, h_c, z_c]) \end{aligned} \tag{7}$$

where trs_{dec_r} and trs_{dec_c} correspond to the forward response decoder and the backward context decoder, respectively. Then, we compute the generic vocabulary token distribution:

$$p(y_t^{(i)}) = softmax(W_v s_t + b_v) \tag{8}$$

where s_t corresponds to r_{ge} or c_{auto}, and $p(y_t^{(i)})$ is the output token distribution at time step t. W_v, b_v are the weights and a bias of the corresponding softmax network.

4.1.4. Emotion Classifier

We introduce an emotion classifier to explicitly detect the emotion from the user utterance. It is trained from the response as well. The emotion classifier is connected with the shared variational layer to achieve emotion consistency between the context and response. We use the CLS embedding h_{c_0} of the encoder output to represent the global memory of the entire context. And we use the cross-entropy as the loss function:

$$\begin{aligned} p_e &= softmax(W_e(h_{c_0} \oplus z_c) + b_e) \\ \mathcal{L}e &= \sum_{i=1}^{n_e} -e_s * \log(p_e) \end{aligned} \tag{9}$$

where W_e, b_e are weights and a bias of the emotion classifier network; e_s is the ground-truth emotion label, n_e is the number of emotion categories.

4.2. Model Optimisation

We describe how to optimize our proposed model in this sub-section. Given the paired datapoint (c, r), the main objective is to optimize the log-likelihood of the joint generation probability $p(c, r)$:

$$\mathcal{L} = \log \int_z p(c, r, z) dz \tag{10}$$

However, this optimization is intractable because of the unknown latent variable z. Inspired by the derivations from Tseng et al. [12], Shen et al. [8], we follow the neural variational inference as introduced in the variational Bayes approach [19], our objective can be achieved by maximizing the evidence lower bound of $\mathcal{L}c, r$ and $\mathcal{L}r, c$:

$$\mathcal{L} \geq \mathcal{L}c, r + \mathcal{L}r, c \tag{11}$$

where $\mathcal{L}c, r$ and $\mathcal{L}r, c$ are the objective function of the forward context model and the backward response model, separately. The former is formulated as:

$$\begin{aligned}\mathcal{L}c, r =& \mathbb{E}_{q_{trs_{enc_c}}(z_c|h_c)} \log p_{trs_{dec_r}}(r_{ge}|z_c, h_c) \\ &+ \mathbb{E}_{q_{trs_{enc_c}}(z_c|h_c)} \log p_{trs_{dec_c}}(c_{auto}|z_c, h_c) \\ &- D_{KL}[q_{trs_{enc_c}}(z_c|h_c)||p(z)]\end{aligned} \tag{12}$$

The first term represents response generation in the forward process; the second term denotes the variational auto-reconstruction of context c_{auto}; the third term means the Kullback-Leibler (KL) divergence between the forward Gaussian posterior $q_{trs_{enc_c}}(z_c|h_c)$ with the prior distribution $p(z)$ of the shared variational layer:

$$\begin{aligned}&D_{KL}[q_{trs_{enc_c}}(z_c|h_c)||p(z)] \\ &= \int_z q_{trs_{enc_c}}(z_c|h_c)[\log q_{trs_{enc_c}}(z_c|h_c) - \log p(z)] dz\end{aligned} \tag{13}$$

where $q_{trs_{enc_c}}(z_c|h_c)$ and $p(z)$ are both the multi-variate standard Gaussian distributions, $p(z)$ denotes previous state of the shared variational layer. Similarly, we can derive a variational optimization objective for the backward response model:

$$\begin{aligned}\mathcal{L}r, c =& \mathbb{E}_{q_{trs_{enc_r}}(z_r|h_r)} \log p_{trs_{dec_c}}(c_{ge}|z_r, h_r) \\ &+ \mathbb{E}_{q_{trs_{enc_r}}(z_r|h_r)} \log p_{trs_{dec_r}}(r_{auto}|z_r, h_r) \\ &- D_{KL}[q_{trs_{enc_r}}(z_r|h_r)||p(z)]\end{aligned} \tag{14}$$

Finally, the entire model is optimized with the sum of $\mathcal{L}c, r$, $\mathcal{L}r, c$ and $\mathcal{L}e$.

4.3. Alternative Retrieval

To alleviate the difficulty of generating appropriate empathetic responses, we incorporate the retrieval process in the testing to serve as a fallback of the generation process as shown in Figure 2. We first compute the emotion distributions of the input context as shown in Equation (9).

Then, we select the corresponding n candidate responses from the pre-defined set based on the predicted emotions, which are taken from the top five candidates of the classification probabilities. We use the same context encoder to encode the selected candidate responses:

$$h_{candi_i} = trs_{enc_c}(candi_i) \tag{15}$$

where $candi_i$ is the i-th selected candidate response, and i ranges from (1 to 5)$\times n$. Then, we compute the similarity score $sim_{i,j}$ between the candidate representation h_{candi_i} and input context h_j:

$$sim_{i,j} = 1 - \arccos(\frac{h_j^\top h_{candi_i}}{\|h_j\| \, \|h_{candi_i}\|})/\pi. \tag{16}$$

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

4.4. Uncertainty Estimator

To select a response from generation or retrieval, we estimate the emotion uncertainty of our DVG model, which is computed by the entropy of the emotion classification probabilities:

$$E_U = \sum_{v=1}^{V} p_e^v \log p_e^v \tag{17}$$

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

$$r = \left\{ \begin{array}{ll} r_{re}, & \text{if} \;\; E_U < u \\ r_{ge}, & \text{if} \;\; E_U \geq u \end{array} \right. \tag{18}$$

5. Experiments on Empathetic Response Generation

5.1. Datasets

5.1.1. Japanese Dataset

We evaluate our approach on the Japanese EmpatheticDialogues [10], which was created by following the original English EmpatheticDialogues [20]. Japanese native speakers were engaged for constructing situation sentences and dialogues. Each dialog contains four utterances by two persons interacted in the form as ‘ABAB.’ For Japanese EmaptheticDialogue, 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. We train and evaluate our model for each turn of *Listener* responding to *Speaker*, and extend *Speaker*’s inquiries one by one from the context history.

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

5.1.2. English Dataset

We also evaluate the effectiveness of our DVG model on the English EmpatheticDialogues [20] with the same split and setting as the Japanese version.

5.2. 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 [21] vectors to initialize the word embeddings. All hyper-parameters of the Transformer model were set the same as in previous work [9]. Following Shen e al. [8] and Tseng et al. [12], we applied KL annealing [22] to alleviate the degeneration issue of the variational network. We used greedy search during inference in the generation process and the maximum decoding step was set to 30.

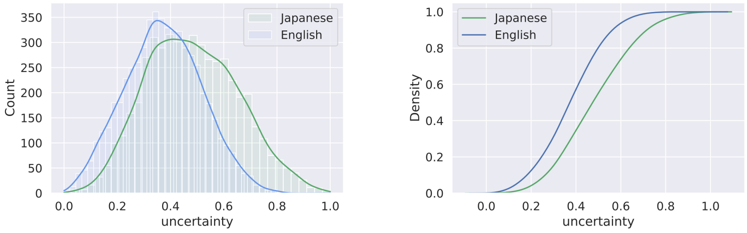
5.3. Comparison Models

For a comprehensive evaluation, we compare our model with other state-of-the-art models.
Transformer [20]: This is a standard Transformer encoder-decoder architecture model. After encoder, it coupled a response decoder and emotion classification.
MoEL [6]: This is an extension of Transformer, which softly combines multiple emotion-specific decoders to a meta decoder to generate an empathetic response.
MIME [7]: This method assumes that empathetic responses often mimic the speaker's emotion and integrates emotion grouping, emotion mimicry, and stochasticity into the emotion mixture for various empathetic responses.
Dual-Emp [8]: This method introduced the dual learning framework, which simultaneously constructs the emotion consensus by a dual-generative model, and also utilizes some external unpaired data. Note that, for a fair comparison, we only compare with this method without using external unpaired data. The major difference from our model is that we also incorporate VAE into the dual generative model with a reconstruction process (e.g. context to context) to enhance the shared layer for better content consistency, in addition to the consistency between the context and response.

5.4. Evaluation Measures

5.4.1. Automatic Metrics

For automatic evaluation, we use the following metrics: (1) PPL (Perplexity) [23] which measures the linguistic complexity of the generated response. (2) BLEU [24] which evaluates the matching of the generated response to the ground truth. We use *multi-bleu.perl* [25] to compute the BLEU scores. (3) EA (Emotion accuracy), which evaluates whether the model correctly recognizes emotion states. There are some similar emotions in the 32 categories. Thus, if the ground truth emotion falls into the top 5 predicted emotions, then we regard the correct prediction. (4) D1/D2 (Distinct-1/ Distinct-2) [26] to evaluate the diversity aspect. (5) BERTScore [27] is a BERT-based evaluation measure for text generation, which focus on lexical semantic similarity between the generated response and the ground truth.



(a) Count distribution

(b) Cumulative distribution

Figure 3. Emotion uncertainty distribution on the validation set of the Japanese and English EmpatheticDialogues dataset.

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

	Uncertainty threshold	Cumulative	Dist-1(%)	Dist-2(%)
Japanese	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
English	0.18	0.1	2.63	8.84
	0.25	0.2	**2.66**	**8.89**
	0.30	0.3	2.67	8.77

5.4.2. 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 [7,8,20]. 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 linguistically correct and readable. Each metric is rated on a scale from 1 to 5.

5.4.3. Human A/B Test

To directly compare the overall performance of our method and others, we also adopt the human A/B test. For two generated responses, one is by our DVG, and the other is from one of the compared models: Transformer, MOEL, MIME, Dual-Emp. Three annotators are asked to choose the better one, or select 'Tie.'

5.5. 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

Table 2. Automatic and human evaluation results of our method and compared models for the Japanese EmpatheticDialogues dataset, bold font denotes the best performances. BERT represents BERTScore. Emp, Rel, Flu are abbreviations of Empathy, Relevance, and Fluency, respectively.

Model	Automatic Evaluation						Human Evaluation		
	PPL ↓	BLEU	EA(%)	D1(%)	D2(%)	BERT(%)	Emp	Rel	Flu
Transformer [20]	20.33	**6.92**	69.25	1.34	5.77	73.21	2.88	2.47	2.89
MoEL [6]	19.49	0.66	68.69	1.36	5.67	73.36	3.15	2.74	2.95
MIME [7]	20.69	0.64	62.46	0.69	2.62	73.08	3.22	2.77	3.24
Dual-Emp [8]	19.23	6.91	71.89	1.11	3.66	73.29	3.22	2.89	**3.30**
DVG (Ours)	**18.32**	6.79	**74.29**	**2.06**	**7.94**	**73.57**	**3.47**	**3.22**	3.24

Table 3. Results of human A/B test for the Japanese EmpatheticDialogues dataset.

DVG (ours) vs.	Win	Loss	Tie
Transformer	42.7%	21.7%	35.7%
MoEL	38.3%	28.7%	33.0%
MIME	38.3%	29.7%	32.0%
Dual-Emp	35.0%	29.0%	36.0%

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 for the Japanese experiments and 0.25 for the English experiments.

5.6. Japanese Dialogue Results and Analysis

5.6.1. Comparison with other Methods

The automatic evaluation results in the left part of Table 2 show that our DVG model outperforms others in the aspects of *emotion accuracy (EA)*, and diversity metrics (*D1* and *D2*). It demonstrates our model's potential to detect emotions more effectively considering both the emotion and content consistency between the context and response, as well as the ability to generate more diverse responses.

Human evaluation results in Table 2 indicate that, among the compared models, our DVG model has the best performance with more than 7.76% and 11.42% improvement on the dimensions of *Empathy* and *Relevance*, respectively. It confirms our model's superiority for suitable emotion and content expression. Especially, compared with *Dual-Emp*, the improvement on the *Relevance* aspect is noteworthy, which indicates that our model can generate responses with contextual appropriateness.

In addition, we conducted pairwise comparisons between *DVG* with the baseline models to directly compare the overall quality of the generated responses. The results of the human A/B test in Table 3 show that the proposed *DVG* is significantly preferred over others by human judges.

5.6.2. Effectiveness of the Alternative Retrieval System

The effectiveness of the alternative retrieval process is shown in Table 4 using the test set. Compared with the generative model *DVG*, both *DVG + Retrieval* (E_u*=0.3)* and *DVG + Retrieval* (E_u*=0.4)* are superior in the automatic evaluation metrics of *Dist-1*, *Dist-2* and human evaluation metircs of *Empathy*, *Relevance*, and *Fluency*. The *BLEU* score is degraded, because there exists a discrepancy between the retrieved statement and the ground-truth response, but the result demonstrates our model gen-

Table 4. Evaluation of the alternative retrieval system for the test set of Japanese EmpatheticDialgues dataset

Model	Automatic Evaluation			Human Evaluation		
	BLEU	Dist-1 (%)	Dist-2 (%)	Empathy	Relevance	Fluency
DVG	**6.79**	2.06	7.94	3.47	3.22	3.24
DVG + Retrieval (E_u=0.3)	5.74	2.18	**8.14**	**3.67**	**3.41**	3.71
DVG + Retrieval (E_u=0.4)	6.23	**2.20**	7.99	3.50	3.29	**3.96**

Table 5. Sample responses generated by our proposed *DVG+ Retrieval* and other models (Original in Japanese and translated to English).

Case 1	Ground truth emotion	誇りに思う Proud
	Predicted emotion	‘誇りに思う’, ‘感謝する’, ‘感動する’, ‘羨ましい’, ‘感傷的になる’ ‘proud’, ‘grateful’, ‘impressed’, ‘jealous’, ‘sentimental’
	Context	SPK1:大学でお世話になった教授の論文が著名な雑誌で発表になったよ。 A paper by one of my professors in college has just been published in a prominent journal. SPK2: それは嬉しいね! That’s nice to hear. SPK1:うん。もう今は何の関係もないけど誇りに思うわ。 Yeah, we don’t have a relationship anymore, but i am proud of it.
	Ground truth response	本当だね。嬉しいし誇らしいね。 It’s true, happy and proud.
	Transformer	すごいね、私もそういうこともあるよね 。 That’s great, I know that happens to me too.
	MoEL	すごいね、頑張って。 That’s great. Good luck.
	MIME	そうなんだね。私も一緒に見てみたいな。 I see. I’d like to see it with you.
	Dual-Emp	そうだね。 Yeah, right.
	DVG (Ours)	そうだね、 **応援しているんだね。** Yeah, right. **You’re cheering them on**.
	Retrieval (Ours)	すばらしいですね。 That’s wonderful.
	DVG+Retrieval (Ours)	すばらしいですね。 That’s wonderful.
Case 2	Ground truth emotion	うしろめたい Guilty
	Predicted emotion	‘さびしい’, ‘うしろめたい’, ‘楽しい’, ‘満足’, ‘感傷的になる’ ‘lonely’, ‘guilty’, ‘happy’, ‘contented’, ‘sentimental’
	Context	専業主婦なのに、一日ゲームにはまっちゃって何にもできなかったよ。 I’m a housewife, but I’ve been stuck in the game all day and haven’t been able to do anything.
	Ground truth response	別にそういう日もあってもいいと思うよ。 I think it’s okay to have such a day.
	Transformer	それはすごいね。 That’s amazing.
	MoEL	それは良かったね。お子さん、何かあったの? That’s good. What happened to your child?
	MIME	そうなんだ。それは楽しみだね。 I see. That’s something to look forward to.
	Dual-Emp	そうなんだ。それは残念だったね。 I see. That was bad.
	DVG (Ours)	そうなんだ。**でも、たまにあるよね。** I see. **But it does happen sometimes**.
	Retrieval (Ours)	気まずいですね。 That’s awkward.
	DVG+Retrieval (Ours)	そうなんだ。でも、たまにあるよね。 I see. But it does happen sometimes.

erates relevant empathetic responses with better diversity and higher scores by human evaluation. It confirms the effectiveness of 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 conclude that emotion uncertainty set to 0.3 is optimal for our model to integrate generation with the retrieval process when experimented on the Japanese EmpatheticDialogue dataset, and the retrieval process has a significant advantage for *Fluency* because the retrieval set is pre-created by a native speaker in advance.

Table 6. Automatic evaluation results of our method and compared models for the English EmpatheticDialogues dataset, bold font denotes the best performances. BERT represents BERTScore.

Model	Automatic Evaluation						Human Evaluation		
	PPL ↓	BLEU	EA (%)	D1 (%)	D2 (%)	BERT (%)	Emp	Rel	Flu
Transformer [20]	37.33	2.61	73.0	2.17	7.78	85.74	3.44	3.07	3.60
MoEL [6]	37.63	2.53	68.13	1.75	6.51	85.91	3.51	3.19	3.46
MIME [7]	36.84	2.51	69.65	1.68	6.21	**85.91**	3.47	3.45	3.66
Dual-Emp [8]	34.52	**2.67**	69.82	1.38	3.96	85.89	**3.59**	3.40	3.63
DVG (ours)	**32.18**	2.61	**75.83**	**2.42**	**8.26**	85.85	3.53	**3.56**	**3.76**

Table 7. Results of human A/B test for the English EmpatheticDialogues dataset.

DVG (ours) vs.	Win	Loss	Tie
Transformer	47.0%	24.7%	28.3%
MoEL	45.3%	34.3%	20.3%
MIME	41.3%	36.3%	22.3%
Dual-Emp	43.3%	25.7%	31.0%

5.6.3. Case Studies

To illustrate the effectiveness of our proposed *DVG*, we present two examples, as shown in Table 5. In the first case, compared with the baselines, our proposed *DVG* generates a response of 'そうだね ∼ *(Yeah, right.)*' to show cognitive understanding of the context and then '応援しているんだね。 *(You are cheering them on.)*' to show the empathy as responding in the perspective of the counterpartner. As the emotion uncertainty of this sample is low, we use the retrieved response 'すばらしいですね。 *(That's fantastic.)*' which is matched to the predicted emotion as the final output.

In the second case, compared with *Transformer*, *MoEL*, and *MIME*, which misunderstand user's situation and emotion, *Dual-Emp* which also utilizes a dual generative model generates an empathetic expression with suitable content. Compared with *Dual-Emp*, our model which additionally utilizes VAE to mimic the process of context/response understanding by reconstruction can generate more context relevant and emotional comfort response, as 'でも、たまにあるよね。 *(But it does happen sometimes.)*' helps relieve the speaker's guilt. In this case, the emotion uncertainty is high, therefore, we adopt the generated response as the final response.

5.7. English Dialogue Results and Analysis

Automatic evaluation in Table 6 and human evaluation results in Table 7 on the English EmpatheticDialoguesdataset indicate our model's superiority over the baselines.

To have an intuitive demonstration, we also list two cases in Table 8. Case 1 shows that *MoEL*, *MIME*, *Dual-Emp* and our *DVG* model can detect the right emotion and show emotional expression, like '*Oh no! I am sorry to hear that.*' However, our *DVG* model generates a response with better empathy, as '*I hope you are okay.*' also expresses concern. In this case, the emotion uncertainty is high, therefore, we adopt the generated response as the final response. Case 2 illustrates our model's ability for context understanding, and the response '*What happened?*' is similar to the ground truth. As the emotion uncertainty of this sample is low, we use the retrieved response *'That's tough.'* which is matched to the predicted emotion as the final output.

Table 8. Case studies of our DVG model and other models on the English EmpatheticDialogues dataset.

Case1	Ground truth emotion	Anger
	Predict emotion	Apprehensive, Embarrassed, Ashamed, Afraid, Guilty
	Context	I just moved to this neighborhood and some dumb criminals shot one of my neighbors and ran into the woods!
	Ground truth response	That's not good. Do you own a gun?
	Transformer	Oh my goodness! i would be too.
	MoEL	Oh no! That is so annoying! I am glad you are able to do it!
	MIME	Oh no! That is terrible. I would have been so scared.
	Dual-Emp	Oh no! I am sorry to hear that. I am sorry to hear that.
	DVG (Ours)	Oh no! I am sorry to hear that. **I hope you are okay**.
	Retrieval (Ours)	I hope everything will be fine.
	DVG+Retrieval (Ours)	Oh no! I am sorry to hear that. I hope you are okay.
Case 2	Ground truth emotion	Sad
	Predict emotion	Sad, Devastated, Lonely, Disappointed, Faithful
	Context	I am so sad because I have read an article about a newborn girl who died because her parents did not believe in medication and doctors.
	Ground truth response	ugh, those articles always get me too ... : (what was wrong with her?
	Transformer	Oh no! I am so sorry to hear that.
	MoEL	I am so sorry to hear that. I hope you are able to find her.
	MIME	That is sad. I am so sorry to hear that.
	Dual-Emp	That is terrible, I am so sorry to hear that.
	DVG (Ours)	I am so sorry to hear that. **What happened?**
	Retrieval (Ours)	That's tough.
	DVG+Retrieval (Ours)	That's tough.

6. Experiments on General Response Generation

In daily-style conversation, an empathetic response is just one kind of conversation reply, while neutral chatting also accounts for a large percentage. Therefore, we further apply our model to a general Japanese dialogue system for the aim of human-robot interaction by incorporating the daily life dataset *PersonaChat* [10] into training. *PersonaChat* focuses on the personality about which the other speaker talks without emotion.

6.1. Dataset

For Japanese PersonaChat, 61,794 utterances are included in the 5,000 collected dialogues. We also train and evaluate the model for each turn of *Listener* responding to *Speaker*, and extend *Speaker*'s inquiries one by one from the context history. We train the model with 33 emotion categories, which consists of 32 emotions used in the EmpatheticDialogue dataset and one additional neutral category for experiments. All the experiment settings are as same as described in Section 5.2.

6.2. Automatic Evaluation

The results in Table 9 show that our model trained with the two datasets does not degrade for each of them. It means our model can generate both empathetic and neutral responses for a general dialogue system. In fact, combining the two datasets contributes to an overall improvement over using a single dataset, even though the topics and emotions are significantly different.

Table 9. Automatic evaluation results when combining *PersonaChat* with *EmpatheticDialogues* dataset. 'Empa' and 'Persona' present EmpatheticDialougues and PersonChat, respectively.

Training	Testset	PPL	BLEU	EA (%)	D1 (%)	D2 (%)	BERT (%)
Empa	Empa	10.97	23.04	70.33	2.08	6.75	79.37
Persona	Persona	16.74	16.38	-	2.36	8.56	73.48
Empa+Persona	Empa+Persona	12.62	19.6	-	2.95	9.42	77.29
	Empa	**10.11**	**24.92**	68.21	**2.43**	**7.97**	79.26
	Persona	**15.44**	16.24	-	**2.92**	**9.37**	**75.50**

6.3. Human Interaction Evaluation

Finally, we evaluate the system in real interaction with human subjects via speech.

6.3.1. System Settings

For each conversational input, if the detected emotion is *neutral*, we adopt a generated response rather than using the retrieval system. Otherwise, as described in Equation (18), if the uncertainty of the detected emotion is smaller than the threshold E_u, we apply the retrieval system.

The context history is important for the system to understand the subject's talking, then generate a consistent and coherent response throughout the conversation. However, the spoken dialogue system is different from the model training which uses clear texts, and it is affected by the errors of the ASR system. Therefore, we make two settings of our *DVG+Retrieval* for comparison: one sets context history to 1 and the other sets to 2.

6.3.2. Reference: Attentive Listening System

For reference, we compared with an attentive listening system [28], which can generate several types of listener responses: backchannels, repeats, elaborating questions, assessments as well as empathetic responses. The system is reported to show comparable performance to the WOZ system in basic skills of attentive listening such as actively listening, encouragement to talk, and focused on the talk.

6.3.3. Experiment Settings

We recruited students from our university to talk with the virtual agent Gene [29], and each subject was given the topic of *'The experience that impressed you most or recently.'* but not constrained to this topic. They are asked to talk with the *Attentive Listening System* [28] and our *DVG+Retrieval* (context=1) and *DVG+Retrieval* (context=2) systems, alternately, based on the given topic. And each conversation lasted 8 minutes. After the conversion, each subject completed the questionnaire on a point ranging from 1 (completely disagree) to 7 (completely agree) for each item, as shown in Table 10. The order of the test system was randomized for each subject.

6.3.4. Results of Human Interactions

Table 10 reports the average score for each question item. The *DVG+Retrieval* system performs overall better than the *Attentive Listening* system. It performs better when the context history is set to 1 than when it is set to 2, but p_{12} value shows that there is no significant difference between the *DVG+Retrieval* (context=1) and *DVG+Retrieval*

Table 10. Average scores on subjective evaluation and t-test results (subjects=21). 'D+R' and 'A' represents our 'DVG+Retrieval' system and 'Attentive Listening' system, respectively. p_1, p_2 and p_{12} mean p-value of the comparation between 'A' and 'D+R (context=1)', 'A' and 'D+R (context=2)', 'D+R (context=1)' and 'D+R (context=2)', separately.

Metric name	Questionnaire Items	A	Context=1		Context=2		
			D+R	p_1	D+R	p_2	p_{12}
Humanness	The system's utterances were human-like and natural.	4.0	4.1	.858	3.8	.658	.498
Cognition	The system understood the talk.	4.0	3.8	.186	3.8	.229	1.00
Emotion	I felt that the system can express various emotions.	4.3	4.2	.800	4.1	.470	.479
Empathy	The system was able to empathize with my experiences.	4.7	4.2	.107	4.0	.017	.217
Personality	I felt that the system has personality.	3.7	3.9	.530	3.6	.812	.322
Agency	I felt that the system was speaking from its own perspective.	2.6	3.6	.002**	3.3	.031*	.507
Topic	I felt that the system had a topic it wanted to discuss.	2.1	3.2	.001**	2.9	.020*	.464
Attentiveness	The system was attentive to me and was actively trying to talk with me.	3.0	4.1	.037*	3.8	.049*	.685
Diversity	The system was able to provide various responses.	3.6	4.3	.061^{+}	4.2	.085^{+}	.893
Engagement	I felt absorbed in the interaction with the system.	3.0	3.7	.079^{+}	3.1	.642	.126
Ease	It was easy to continue a conversation with the system.	2.9	3.5	.094^{+}	3.1	.448	.185
Enjoyability	I enjoyed speaking with the system.	3.3	3.6	.425	3.1	.463	.046*
Talk again	I want to talk with the system again.	3.1	3.4	.464	3.0	.825	.249

$(**p < .01, *p < .05, +p < .1)$

(context=2) system. We observe subjects often switch emotions or topics within the conversation, in this case, the *DVG+Retrieval* (context=2) system tends to generate inappropriate responses because both emotion and topic are consistent within each conversation in our training datasets.

Specifically, the *DVG+Retrieval* (context=1) system achieved a significantly better score than the *Attentive Listening* system for the evaluation of *Agency*, *Topic*, *Attentiveness*, *Diversity*, *Engagement*, *Ease*. This indicates that *DVG+Retrieval* (context=1) system can enrich the conversation with diverse topics and responses as well as be actively attentive to users. No significant difference was observed between the two systems under the evaluation of *Humanness*, *Cognition*, *Emotion*, *Empathy*, *Personality*, *Enjoyability* and *Talk again*. The *Attentive listening* system focused on keyword detection of the input, then produced template-based responses. Thus, it tends to produce safe but proto-typical responses. On the other hand, the proposed system can generate more diverse responses depending on the context. But it is prone to ASR errors and often results in irrelevant responses.

To further examine the effect of the retrieval model in the *DVG+Retrieval* (context=1) system, we calculated the ratio between the retrieved responses against all responses. When we pick up the sessions when the retrieval ratio was larger than 10%, the *DVG+Retrieval* (context=1) system was preferred by humans over the *Attentive Listening* system in all of the subjective evaluations. This suggests that when confident emotion recognition is performed, the system works much better.

6.3.5. Future Perspective

To take advantage of both *Attentive Listening* and *DVG+Retrieval*, we plan to build a hybrid system combining both systems. Specifically, we take the *Retrieval* system as the first priority to produce an emotion-specific response when the system is confident about the recognized emotion. *Attentive Listening* system is in the second priority if it generates a response in the type of 'Repeat' or 'Questions', which is safe and relevant to the context. In other cases, we can turn to the *DVG* system, which can enrich the conversation with diverse topics and responses.

7. Conclusions

In this paper, we have proposed the DVG model for empathetic response generation. Our DVG model can efficiently capture the mutual characteristics of the content and emotion consistency between the context and the response. Evaluations on both Japanese and English *EmpatheticDialogues* datasets demonstrate our model's superiority in generating empathetic responses with contextual and emotional appropriateness. In addition to the DVG model, we proposed an auxiliary retrieval system to improve empathetic response generation. We further extended our model's potential in generating both empathetic and general responses, and implemented in the human-robot interaction dialog system.

References

[1] Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S. & An, L. Understanding and predicting empathic behavior in counseling therapy. *ACL*. pp. 1426-1435 (2017)

[2] Fitzpatrick, K., Darcy, A. & Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*. **4**, e7785 (2017)

[3] Prendinger, H. & Ishizuka, M. THE EMPATHIC COMPANION: A CHARACTER-BASED INTERFACE THAT ADDRESSES USERS'AFFECTIVE STATES. *Applied Artificial Intelligence*. **19**, 267-285 (2005)

[4] Wang, L., Wang, D., Tian, F., Peng, Z., Fan, X., Zhang, Z., Yu, M., Ma, X. & Wang, H. CASS: Towards building a social-support chatbot for online health community. *ACM On Human-Computer Interaction*. **5**, 1-31 (2021)

[5] Davis, M. Measuring individual differences in empathy: evidence for a multidimensional approach.. *Journal Of Personality And Social Psychology*. **44**, 113 (1983)

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

[7] 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)

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

[10] 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)

[11] Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. & Choi, Y. Atomic: An atlas of machine commonsense for if-then reasoning. *AAAI.* **33**, 3027-3035 (2019)

[12] Tseng, B., Cheng, J., Fang, Y. & Vandyke, D. A Generative Model for Joint Natural Language Understanding and Generation. *ACL.* (2020)

[13] Cui, S., Lian, R., Di Jiang, Y., Bao, S. & Jiang, Y. DAL: Dual Adversarial Learning for Dialogue Generation. *NAACL Workshop On Methods For Optimizing And Evaluating Neural Language Generation.* pp. 11-20 (2019)

[14] Hu, X., Wang, R., Zhou, D. & Xiong, Y. Neural Topic Modeling with Cycle-Consistent Adversarial Training. *(EMNLP).* pp. 9018-9030 (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. *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] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances In Neural Information Processing Systems.* **30** (2017)

[18] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805.* (2018)

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

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

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

[22] Bowman, S., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R. & Bengio, S. Generating Sentences from a Continuous Space. *CoNLL.* pp. 10-21 (2016)

[23] Vinyals, O. & Le, Q. A neural conversational model. *ArXiv Preprint ArXiv:1506.05869.* (2015)

[24] Papineni, K., Roukos, S., Ward, T. & Zhu, W. Bleu: a method for automatic evaluation of machine translation. *ACL.* pp. 311-318 (2002)

[25] Britz, D., Goldie, A., Luong, M. & Le, Q. Massive Exploration of Neural Machine Translation Architectures. *EMNLP.* pp. 1442-1451 (2017)

[26] 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)

[27] Zhang, T., Kishore, V., Wu, F., Weinberger, K. & Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *ICLR.* (2019)

[28] 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. *SIGDIAL.* pp. 118-127 (2020)

[29] Lee, A. & Ishiguro, H. Development of CG-based Embodied Dialogue Agents and System with Conversational Reality for Avatar-Symbiotic Research. *SIG-SLUD, JSAI.* (2022)