ER-Chat: A Text-to-Text Open-Domain Dialogue Framework for Emotion Regulation

Shin Katayama, Shunsuke Aoki, Takuro Yonezawa, Tadashi Okoshi, Jin Nakazawa, and Nobuo Kawaguchi

Abstract—Emotions are essential for constructing social relationships between humans and interactive systems. Although emotional and empathetic dialogue generation methods have been proposed for dialogue systems, appropriate dialogue involves not only mirroring emotions and always being empathetic but also complex factors such as context. This paper proposes Emotion Regulation Chat (ER-Chat) as an end-to-end dialogue framework for emotion regulation. Emotion regulation is concerned with actions to approach appropriate emotional states. Learning appropriate emotion and intent when responding on the basis of the context of the dialogue enables the generation of more human-like dialogue. We conducted automatic and human evaluations to demonstrate the superiority of ER-Chat over the baseline system. The results show that inclusion of emotion and intent prediction mechanisms enable generation of dialogues with greater fluency, diversity, emotion awareness, and emotion appropriateness, which are greatly preferred by humans.

Index Terms—Emotion Regulation, Affective Computing, Dialogue Generation

1 Introduction

Open-domain dialogue systems, such as chit chat, aim to establish long-term connections with users by satisfying the human need for communication, affection, and social belonging [1]. To ensure more practical and human-like dialogue systems, extensive research has been conducted on affective computing [2]. Previous research has shown that emotional dialogue systems can improve user satisfaction [3] and increase positive interactions [4]. Dialogue systems can be considered as digital partners that handle emotions appropriately and apply them to mental healthcare and counseling.

Advancements in the development of natural language processing using deep learning have enabled the proposal of neural dialogue generation methods aimed at realizing human-like and natural dialogues. In affective computing, emotional dialogue generation [5], [6], [7] and empathetic dialogue generation [8], [9] have been performed by applying various neural dialogue generation methods. However, emotional dialogue generation methods do not allow seamless dialogues because they require manual selection of the emotion labels. In addition, empathetic dialogue generation focuses on recognizing or copying the speaker's emotions and does not discuss the listener's appropriate responses. We argue that recognizing or copying emotions is insufficient for human-like dialogue because the listener's appropriate response involves several complex factors beyond emotion alone.

In psychology, emotional regulation to reduce negative emotions and approach appropriate emotional states is considered essential to maintain mental health, social functioning, and physical health [10], [11]. Furthermore, interpersonal emotion regulation allows accurate judgment and contributes to the development of social relationships [12]. To achieve interpersonal emotion regulation, it is essential to recognize situations and implement appropriate actions [13]. Therefore, a dialogue system that aids emotion regulation must recognize the speaker's potential context and provide appropriate dialogue responses. This action involves a natural series of activities performed by humans in human-to-human communication.

Fig. 1: Example of the EmpatheticDialogue [8] dataset based on an emotional scenario.

Figure 1 shows an example of the EmpatheticDialogue [8] dataset based on an emotional scenario. Given the input text of the speaker *"I've been feeling a bit blue and overwhelmed lately"*, the listener responds with *"You should take a vacation that will help you relax and see new stuff"*. As the speaker's input text can capture scenarios such as sad or lonely, if the empathetic dialogue system were to respond, it is expected to give a sad response such as *"I'm sorry to hear that"*. However, the actual data attempt to regulate the speaker's emotion by responding with a caring emotion, indicating the intent to make a suggestion. It is possible to regulate emotions by asking other questions and engaging

- *Shin Katayama is with the Graduate School of Engineering Nagoya University, Aichi, Japan.*
  Email: shinsan@ucl.nuee.nagoya-u.ac.jp
- *Shunsuke Aoki is with the National Institute of Informatics, Tokyo, Japan.*
- *Tadashi Okoshi and Jin Nakazawa are with Faculty of Environment and Information Studies, Keio University, Kanagawa, Japan.*
- *Takuro Yonezawa and Nobuo Kawaguchi are with Graduate School of Engineering, Nagoya University, Aichi, Japan.*

Manuscript received April 19, 2005; revised August 26, 2015.

in this context, so empathy alone is insufficient. Thus, although appropriate dialogue is a difficult task without a correct answer, it is possible to achieve a more human-like and satisfactory dialogue by following appropriate emotion and intent.

This paper proposes Emotion Regulation Chat (**ER-Chat**) as an open-domain dialogue framework for emotion regulation. ER-Chat is responsible for dialogue generation using T5 [14], a pre-trained language model. The classifier is used to automatically assign annotated emotion and intent labels to sentences in the EmpatheticDialogue dataset, and human-like dialogue is learned by combining the listener's predicted emotion and intent losses according to context. We believe that dialogue systems can improve user interaction experiences by learning how to respond with appropriate emotions and intent for emotion regulation. In the experiments, we conducted automatic and human evaluations for the quality of the generated dialogues. The automatic evaluation is comparable to the state-of-the-art method, and human evaluation shows that the proposed method is superior, indicating that the proposed method generates sentences that satisfy the user in terms of emotional awareness and emotional appropriateness while maintaining the quality of the responses, including fluency and diversity.

The contributions of this paper are as follows.

1) We propose ER-Chat, a text-to-text dialogue framework for chit chat that aims at emotion regulation by predicting appropriate emotions and intentions during responses.
2) Automatic evaluation using dialogue generation metrics and human evaluation by a total of 200 subjects confirm that our proposed method is effective.
3) We conduct case studies to discuss dialogue generation for emotion regulation and suggested directions for future research.

2 Related Research

2.1 Neural Dialogue Generation

Although current dialogue systems cannot perfectly produce smooth and precise dialogue as humans do, advances in machine learning and deep learning technologies have made it possible to achieve human-like dialogue under limited conditions [15]. Response methods for open-domain dialogues have traditionally been rule-based or search-based, but in recent years, neural network-based dialogue generation methods using Seq2seq [16] or Transformer [17] have become mainstream. Furthermore, large-scale pre-trained language models such as BERT [18], GPT [19], and T5 [14] are widespread. These models are based on Transformer, which learns the amount of semantic knowledge from a large corpus. They can be applied to specific NLP tasks with only a small corpus by fine-tuning. In open-domain dialogues, the goal is to realize more human-like dialogue systems by using persona information [20], [21] or characteristics [22] to generate controllable dialogue.

2.2 Emotional/Empathetic Dialogue Generation

Emotional dialogue generation based on neural dialogue generation has been actively studied as an approach to improve dialogue performance. Zhou et al. [5] proposed Emotional Chatting Machine (ECM), the first approach that enables large-scale emotional expression using deep learning. Since then, diverse research has established methods for generating end-to-end emotional responses using neural networks [6], [7], [23], [24], [25], including VAE based method [26], multi emotion label method [27], and multi-turn dialogue method [28]. Shen et al. [29] proposed Curriculum Dual Learning, which extends emotion-controllable response generation to a dual task to generate emotional responses and emotional queries alternatively. However, these methods focus on controlling the emotional content of the textual response through a manually specified emotion label.

In order to approach real-world dialogue scenarios, an empathetic dialogue generation method [30], [31], [32], such as recognizing emotions in input sentences and generating responses [33], dialogue response by eliciting positive emotions [34], has been proposed. Liang et al. [35] proposed a response generation method based on appropriate emotions, taking into account the personality of the speaker. Li et al. [9] proposed to leverage multitype knowledge to understand and express emotions explicitly for empathetic dialogue generation. Ma et al. [36] proposed a control unit for the generation of emotional dialogue to deal with emotion drift, where the emotion at the time of input is different from the emotion at the time of response. Wei et al. [37] proposed more intelligent responses that adequately express emotions by simultaneously encoding the semantics and emotions of posts. These studies aimed to empathize and allow for dialogue responses, but empathy is not the only appropriate response. Emotional regulation in interpersonal communication [38], [39], [40] has been actively discussed in psychology, and models for interpersonal emotion regulation have been proposed based on multiple and complex contexts, not only empathy. Therefore, in this study, we implemented a data-driven dialogue framework that considers context and responds with appropriate emotion and intent.

3 ER-Chat

We propose ER-Chat, which is an open-domain dialogue framework for emotion regulation. As described above, Transformer-based pre-trained models have yielded remarkable results in various NLP tasks. ER-Chat is based on T5 (Text-To-Text Transfer Transformer) [14]. T5 is a pre-trained model based on a Transformer that solves the text-to-text framework and has achieved state-of-the-art results in various language generation tasks, including translation, question-and-answer, and summarization. Using this architecture, dialogue generation can be realized for emotion regulation by fine-tuning the multitasking of the emotion and intent prediction mechanism. Figure 2 shows an overview diagram of the proposed framework.

3.1 Dataset

Several dialogue corpora have been published for open-domain neural dialogue generation. Rashkin et al. [8] published the EmpatheticDialogues dataset to train and evaluate empathetic dialogue systems. This dataset has 24,850

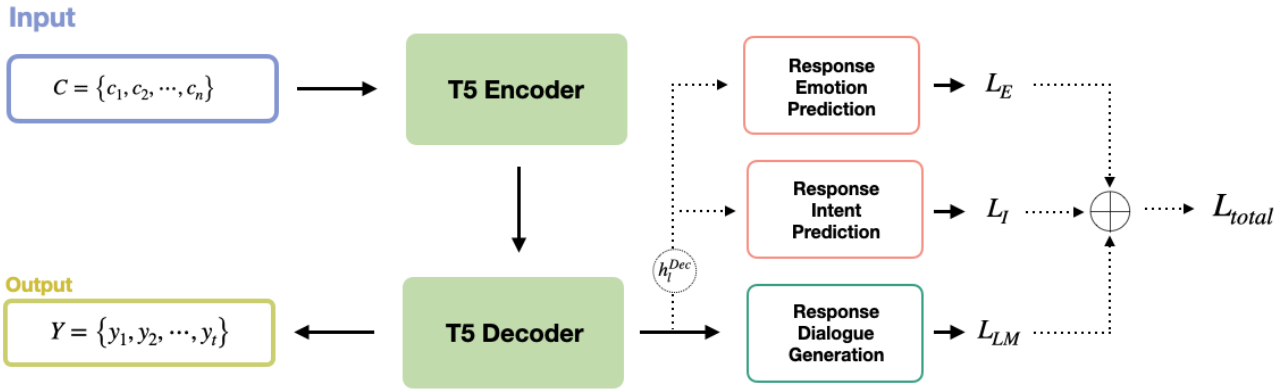

Fig. 2: This figure is a framework structure diagram of our proposed method, ER-chat.

empathetic dialogues based on 32 equally distributed emotions. However, while this dataset is one of the best corpora for the generation of emotional dialogue, it only contains emotion labels for the entire context. Emotion regulation is shown in Figure 1, where listeners interact with different emotions and intents in their sentences. Therefore, we augmented this dataset with emotion and intent labels. To minimize human effort, we built a classifier and performed automatic annotations. Each sentence thus contains two labels, namely emotion and intent. The details of the labels and heat maps for each emotion and intent are shown in Figure 3.

Emotion Label

Ekman's Universal Emotions Theory [41] is often used for emotion labeling. However, in this study, we implemented a classifier based on GoEmotions [42] for more detailed emotion labeling. GoEmotions has 27 categories and neutral labels based on 58,000 comments sourced from Reddit. However, due to the disproportionate number of emotion labels in the dataset, we selected nine emotions and neutral when using the hierarchical clustering proposed by Demszky et al. [42]. Based on the official implementation[1] of the pre-trained BERT base model, ten emotion labels (*Joy, Caring, Admiration, Gratitude, Approval, Surprise, Fear, Sadness, Anger, Neutral*) were assigned to each sentence in the EmpatheticDialogue dataset. The classification performance was reasonable with an accuracy of 65.1 and a macro-F1 score of 65.2.

Intent Label

Welivita et al. [43] proposed an intent-based taxonomy in human social conversations, for which datasets are publicly available. They manually annotated 15 different intents for the EmpatheticDialogue dataset, eventually obtaining 8 high-frequency intents (*Questioning, Acknowledging, Consoling, Agreeing, Encouraging, Sympathizing, Suggesting, Wishing*) and (*Neutral*. We trained a BERT-based classifier and achieved automatic intent labeling. The classification performance was also reasonable with an accuracy of 85.4 and a macro-F1 score of 85.6.

1. https://github.com/google-research/google-research/tree/master/goemotions

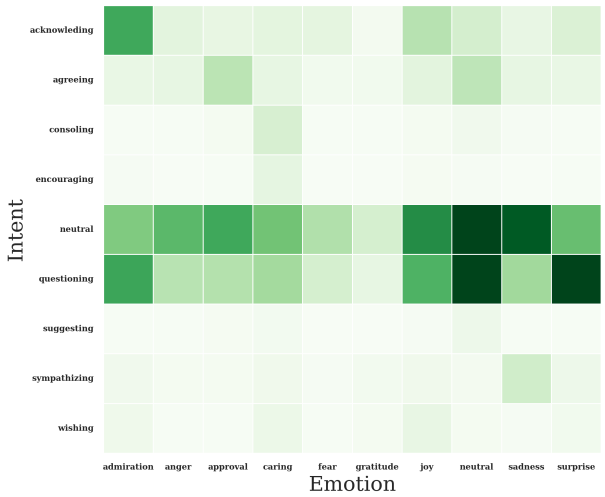

Fig. 3: Heatmap to label the emotion and intent of sentences in the EmpatheticDialogue dataset.

3.2 Methodology

Our proposed approach is a text-to-text dialogue framework for emotion regulation. ER-Chat is a multitasking extension of the Transformer-based T5 model with added modules for emotion and intent predictions. T5 is a deep neural network with high flexibility and performance that allows one pre-trained model to be transfer-trained for diverse tasks by redefinition of the text-to-text problem format. Learning based on real-world knowledge based on T5 architecture enables generation of responses with appropriate emotions and intent while inheriting human-like response generation capabilities. The data for the training process are formulated as follows:

$$\mathcal{D} = \{C, R, E, I\} \tag{1}$$

where $C = \{c_1, c_2, \ldots, c_n\}$ is a dialogue context; the content of C_i consists of $C_i = \{x_1, x_2, \ldots, x_t\}$, with the odd numbered content being speakers and even numbered ones being listeners; $R = \{r_1, r_2, \ldots, r_n\}$ is a gold standard response; E is a response emotion label; and I is the response intent label for prediction.

The proposed method consists of three modules, namely response dialogue generation, response emotion prediction, and response intent prediction. The context C is first passed through T5 encoder and T5 decoder, before being applied to the last hidden layer in the decoder h_l^{Dec} and to the three different modules. During inference, response text $Y = \{y_1, y_2, \ldots, y_n\}$ is generated through these networks.

- **Response dialogue generation**: To optimize response generation, we use a contextual representation of the gold standard response and model the response tokens using cross-entropy loss. This loss is referred to as L_{LM}.
- **Response emotion prediction**: To optimize the response emotions, h_l^{Dec} is passed through two linear layers and an activation function, before applying response emotion prediction using cross-entropy loss. This loss is referred to as L_E.
- **Response intent prediction**: In addition to response emotion prediction, response intent prediction is performed using the cross-entropy loss from the annotated response intent labels. This loss is referred to as L_I.

The final loss function of the fine-tuning step is a weighted sum of the aforementioned three losses.

$$L_{total} = L_{LM} + \lambda(L_E + L_I) \tag{2}$$

where λ is the hyperparameters of the auxiliary losses. Minimizing this total loss can optimize dialogue generation, as well as response emotion and intent predictions.

4 Experiment

4.1 Implementation Details

We implemented all the models on Pytorch, a deep-learning framework. The proposed method used T5-base model with 220 million trainable parameters from HuggingFace[2]. The model has 12 layers of feedforward networks, 768 hidden states, and 12 heads. The emotion and intent prediction heads are in a 300-dimensional space. During training, the learning rate was 1e-5, batch size was 8, and optimizer was Adam [44]. The auxiliary losses λ was set to 1.0. During inference, we used the top-p [45] and top-k [46] sampling methods for the decoding algorithms, where top-k was 20 and top-p was 0.9. We set the max length of the generated response equal to 40 and the min length equal to 4. Our models were then fine-tuned for batch sizes of 20 epochs on a single NVIDIA TITAN V GPU. To maintain an optimal model, an early stopping was applied when the best loss exceeded for three consecutive epochs during training.

4.2 Baseline

ER-Chat performs emotion regulation to optimize emotion and intent in dialogues. Therefore, we conducted the following baseline and comparative evaluation experiments to assert the effectiveness of the proposed method. Models that manually select emotion labels, such as ECM [5], are not employed in the baseline because seamless dialogue is not possible.

- **MoEL [30]:** This Transformer-based model softly combines response expressions from different Transformer decoders. Each decoder is optimized to focus on a specific emotion, and emotion recognition is used to achieve empathetic dialogue.
- **EmpDG [9]:** This model is also a Transformer-based, multitype knowledge-aware empathetic dialogue generation framework. Here, common-sense knowledge and affective vocabulary are used to enrich dialogue utterances using both coarse-grained dialogue-level and fine-grained token-level emotions.
- **T5-base:** This is a pre-trained language model in the Colossal Clean Crawled Corpus (C4) dataset. T5-base model is fine-tuned for the EmpatheticDialogue dataset to learn dialogue generation tasks.

The following ablation tests were also conducted for the automatic evaluation to clarify the impacts of the emotion and intent prediction heads on the proposed method.

- **w/o Intent:** This is a fine-tuned model that considers only emotion prediction loss without intent loss.
- **w/o Emotion:** This is a fine-tuned model that considers only intent prediction loss without emotion loss.

4.3 Metrics

4.3.1 Automatic Evaluation

To evaluate the quality of the response text generated by the proposed method, we conducted an automatic evaluation experiment using evaluation metrics for dialogue generation. Evaluation metrics such as BLEU [47] scores are used during translation to evaluate the document generation models; however, these metrics have been shown to be ineffective in terms of their correlations with human judgment in dialogue generation [48]. Therefore, the following indicators are used for the automatic evaluations.

- **Perplexity:** Perplexity defined as raised to the power of cross-entropy is retained as a reference. Perplexity measures how well the model predicts the data in the test set; in other words, it indicates the accuracy of anticipation of what the users may say next. When interpreting perplexity scores, it should be noted that lower values are better and that the theoretical minimum is one.
- **Distinct-N [49]:** Distinct-N measures and evaluates the diversity of the model output. It can calculate the degree of lexical diversity by calculating the number of unigrams (Distinct-1) and bigrams (Distinct-2) in the generated responses. We evaluated the consistency and diversity of the output sentences in the proposed method using these two metrics.
- **BERTScore [50]:** BERTScore is used as a measure of the fluency of the model output. The tokens in the generated and gold standard sentences are each converted to vector representations using BERT. The similarity between the generated and gold standard

2. https://huggingface.co/t5-base

TABLE 1: Example of generated text using the baseline and ER-Chat by EmpatheticDialogue dataset context.

Context	Model	Generated Text
Speaker: I'm leaving tomorrow to go on vacation! **Listner**: Hey that's awesome! Where ya goin? **Speaker**: We are going to Daytona and Orlando for 10 days.	MoEL	That sounds exciting ! I hope you are happy for you !
	EmpDG	I have never had a couple of nostalgia.
	T5-base	That sounds like a lot of fun!
	ER-Chat	Those are awesome! I'm going to Florida.
Speaker: I have always more trust on my girlfriend.	MoEL	That is good. What are you going to do?
	EmpDG	I am sure you are feeling better now.
	T5-base	Why?
	ER-Chat	I agree. That's a good thing, but trust can be hard to find.
Speaker: My favorite team blew a big lead in the playoffs. It was like drinking sour milk. **Listner:** im a browns fan i feel that pain **Speaker:** I hate when teams choke. It makes me sick.	MoEL	I agree, I am sure you are doing it !
	EmpDG	I am so sorry to hear that. It is hard to be a pain.
	T5-base	I would be so mad!
	ER-Chat	yea I don't blame you for it.
Speaker: I can't help feel stressed about money.	MoEL	You should have a good relationship.
	EmpDG	oh no ! that is a shame.
	T5-base	Why is that?
	ER-Chat	Why? Is it the money you have that keeps you busy?

sentences is determined by the weighted cosine similarity using these distributed representations. We use F1-score for metrics that is the harmonic mean of the precision and recall.

4.3.2 Human Evaluation

To evaluate the response quality of the proposed method, human evaluations with annotations of the subjects was performed. To ensure quality, we conducted two types of experiments, namely annotation evaluation based on the Likert scale using five evaluation indices and A/B test evaluation. As text data for evaluation, 100 dialogue contexts corresponding to the test data were extracted from the dataset. Of these, half were used for the Likert scale evaluation experiment and the remaining half were used for the A/B test. Two hundred subjects were recruited using Amazon Mechanical Turk[3], a crowdsourcing service. The subjects were Master users and therefore experts at using the Amazon Mechanical Turk; all of them live in the U.S. and are fluent in English. The age of the subjects ranged from 20 to 60 years; there were 114 males and 84 females, and two subjects did not prefer to tell us their genders. A reward of ten dollars was given per person, and the response time was approximately one hour. Examples of generated text used during the manual evaluations are shown in Table 1.

Likert Scale Test

The five evaluation measures for the response quality were Fluency, Relevance, and Empathy, which were used in the human evaluations of the baseline method, in addition to Emotion Awareness and Emotion Appropriateness, which are indices assessed to clarify whether the dialogue generation was directed toward emotion regulation. Responses were generated by the proposed method and three baseline methods for each of the 50 dialogue context samples. To ensure fairness, each response was presented in random order, and the subjects annotated all responses on a five-point Likert scale (1 = very bad, 2 = bad, 3 = neutral, 4 = good, 5 = very good). A sample screen of the Likert scale test is shown in Figure 4.

- **Fluency:** *How linguistically intelligible is the response?*
- **Relevance:** *How relevant is the response to the context and user?*
- **Empathy:** *How empathetic is the response?*
- **Emotion Awareness:** *How aware is the Bot regarding the user's emotion?*
- **Emotion Appropriateness:** *How appropriate is the response in dealing with emotions?*

A/B Test

To further solidify the human evaluations, the A/B test was conducted to compare the dialogues with their contexts. This test demonstrates the superiority of the proposed method by directly comparing the responses generated by ER-Chat with those of the baseline method. Fifty annotators were asked to choose among *1. ER-Chat is better*, *2. Both are equal*, and *3. (MoEL, EmpDG, T5) is better* for a sample of 50 randomly ordered contexts and responses.

4.4 Results

4.4.1 Automatic Evaluation

The left side of Table 2 details the results of the automatic evaluations. Compared to the baseline method, the output text in ER-Chat had higher Distinct-2 and BERTScore values. However, for Perplexity and Distinct-1, the highest score was obtained in the ablation test or baseline. As the BERTScore is an index used to evaluate similarity between the generated sentences and BERTScore, the proposed method was observed to allow for more emotional responses than the existing methods by regulating emotions against the gold standard response. Further, Distinct-N is indices that evaluate the diversity of the generated dialogues; in terms of these indices, the performance of the proposed method was definitely superior, indicating that it allows more consistent and diverse outputs. Since these differences were negligible, it cannot be concluded that the

3. https://www.mturk.com/

TABLE 2: Result of Auto/Human evaluation

	Automatic Evaluation				Human Evaluation				
	Perplexity	Distinct-1	Distinct-2	BERTScore (F1)	Fluency	Relevance	Empathy	Emotion Awareness	Emotion Appropriateness
MoEL	33.85	0.0122	0.1287	0.8484	3.309	2.076	2.390	2.135	2.043
EmpDG	34.31	0.0128	0.1173	0.8507	3.355	2.172	2.633	2.309	2.204
T5-base	12.483	**0.0310**	0.3153	0.8651	4.402	3.765	3.426	3.357	3.336
ER-Chat	12.747	0.0295	**0.3168**	**0.8654**	**4.482**	**3.788**	**3.545**	**3.455**	**3.441**
w/o Intent	12.518	0.0284	0.3015	0.8634	-	-	-	-	-
w/o Emotion	**12.482**	0.0286	0.2991	0.8631	-	-	-	-	-

TABLE 3: Result of A/B test in Human evaluation

Models	Win	Tie	Lose
ER-Chat VS MoEL	**65.30%**	23.50%	11.20%
ER-Chat VS EmpDG	**72.08%**	18.16%	9.76%
ER-Chat VS T5-base	**40.04%**	25.44%	34.52%

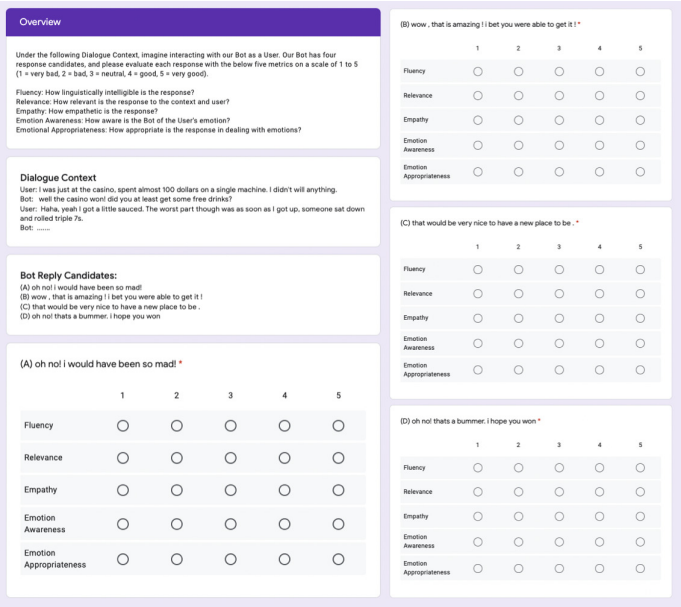

Fig. 4: Actual screenshot of a Likert scale experiment in human evaluation.

proposed method significantly outperformed the baseline SoTA method, but we showed comparable results to the SoTA method.

4.4.2 Human Evaluation

The right side of Table 2 shows the results of the human evaluations. All indicators, namely Fluency, Satisfaction, Relevance, Emotion Awareness, and Emotion Appropriateness, show that our proposed method outperforms the baseline method. In particular, because ER-Chat and T5-base use pre-trained language models, they significantly outperform the Transformer-based MoEL and EmpDG for all metrics. Similarly, the two emotional indices, namely Emotion Awareness, and Emotion Appropriateness, were above those of the baseline, thus confirming that the output text of the proposed method included appropriate emotions. Table 3 shows the results of the A/B test performed on the results of the automatic and human evaluations. When subjects were asked to annotate which approach was better, the results showed that the proposed approach was much better than the baseline method, similar to the results of the Likert scale test.

We conducted repeated measures ANOVA tests using the proposed method and the three baseline assessments based on the aforementioned results. The results showed that five metrics: Fluency, Satisfaction, Relevance, Emotion Awareness, and Emotion Appropriateness, were significantly different ($p < 0.01$) for all five indicators. Therefore, multiple comparisons were conducted using Tukey's test as a post-hoc comparison, and significant differences ($p < 0.01$) were found for all five indicators for ER-Chat, MoEL, and EmpDG. Furthermore, no significant differences were found in multiple comparisons for ER-Chat and T5-base. The reason for this may be that the pre-trained model was used, which could not significantly outperform the proposed method due to dependencies such as vocabulary usage.

Table 3 shows the results of the A / B test performed based on the results of automatic and human evaluations. When subjects were asked to annotate which was better, the results were superior to those of the baseline method, as were the results of the Likert scale. Slightly better results were also obtained in comparison with those from T5-base. We conducted a paired t-test for the results of this A / B test, and found a significant difference ($p < 0.01$) for ER-Chat vs. MoEL and ER-Chat vs. EmpDG. Although there were no significant differences observed for ER-Chat vs. T5-base, the proposed method, outperformed all the baseline methods in the percentage of A / B test selections.

4.5 Case Study

We conducted a case study to identify the context important for emotion regulation. Table 4 shows the input contexts and emotion scenarios from the original dataset sampled from EmpatheticDialogue, ER-Chat responses to these, and the gold standard responses, which denote the actual response text in the dataset.

Case 1 describes a scenario in which the speaker is feeling lonely when going to the movies alone. ER-Chat response in this situation suggests inviting a friend. This response scenario is more like a human conversation, something that cannot be generated by lonely scenario detection and empathy alone. Thus, this is one example of how ER-Chat can achieve emotion regulation.

TABLE 4: Examples of context and response in the EmpatheticDialogue dataset for a case study.

Case 1	Scenario	**Lonely**
	Context	**Speaker:** I was going to go to the movies with a bunch of friends. But they all bailed on me and now I have to go alone. **Listener:** If the movie is worth it you will never miss them. **Speaker:** I hope so, but I am really missing them right now. I didn't want to go alone.
	ER-Chat	**You should go with a group of friends if you don't mind me asking. Its a good feeling.**
	Gold	I understand. Its tough these days.
Case 2	Scenario	**Grateful**
	Context	**Speaker:** Im so grateful for my family due to an incident with my friend. **Listener:** Oh really? What happened with your family and friend then? **Speaker:** My friend lost her parents and siblings in a fire and their family was very close to mine. So seeing her go through the grief made me realize how thankful I should be.
	ER-Chat	**Wow, that is so sad to hear. I have a friend who is a firefighter and I am grateful for her.**
	Gold	Oh wow, that's awful. Stuff like that really teaches you to appreciate what you've got and to not take it for granted.
Case 3	Scenario	**Annoyed**
	Context	**Speaker:** I have a friend that has a child that is unruly and does bad things. One day at our home, he grabbed a game and sat on it, instantly breaking it! It was expensive and this really irked me, especially because she doesn't discipline him when he does bad things.
	ER-Chat	**Oh my god. That sucks! Its a hard situation for him to be in.**
	Gold	oh man. i'm all about discipline! I don't like spoiled bratty kids.
Case 4	Scenario	**Proud**
	Context	**Speaker:** i'm so proud of my son! He learned a couple of words this week. He now knows how to ay please **Listener:** That's wonderful, they say children have a certain gift for picking up languages. **Speaker:** he actually speaks 2 languages. I teach him english and russian
	ER-Chat	**That's amazing, I bet he is a very smart kid.**
	Gold	Wow! That's impressive and a lasting skill for your child to develop.

Case 2 describes a scenario in which the speaker is grateful to his family. ER-Chat response not only owns the grateful emotion but also uses words such as *firefighter* to capture the context. Thus, the response can reliably take context into account even in multi-turn dialogues. The response is as empathetic as Gold's, and natural responses could be generated.

Case 3 describes a scenario in which a person is angry with his or her child. ER-Chat's response is similar to Gold's, with anger emotion and agreeing intent. In this context, dialogue systems in which the response is always positive will appease or calm the speaker. However, ER-Chat facilitates dialogue by responding with the same emotion as the speaker. Thus, ER-Chat generates with appropriate emotion and intent, considering the context.

Case 4 describes a scenario in which a person is proud of his son. ER-Chat responds with the same surprise emotion and acknowledging intent as Gold and generates appropriate dialogue based on context. As mentioned above, there is no correct answer for generating appropriate dialogue. As a result, it often differs from Gold's response as well, but most of the responses generated by ER-Chat are more natural and are one of appropriate responses. Therefore, it is essential to use the information on emotions and intents to generate more human-like dialogues.

5 Discussion for Future Works

Emotion and Intent Limitation

In this research, we used an emotion taxonomy of GoEmotions [42]. However, it must be noted that, other methods, such as Russell's circumplex model [51], which treats emotions as continuous values rather than independent categories [52], have been proposed. In addition, we used existing methods for the intent taxonomy [43]; however, this approach is limited as it cannot perform exact labeling. Furthermore, our approach does not entirely remove noise because of the automatic labeling of emotions and intent using BERT. Therefore, expressions of emotion and intention can be handled in a more detailed manner by using vector representations or continuous values.

Context Limitations

It is necessary to estimate the appropriate emotion for contextual information such as user behavior and sentence content for emotion regulation. Therefore, if this dialogue system is to be implemented as a real application, the long-term context, such as dialogue history, should be considered. In addition, the generation of open-domain dialogue is a difficult task with no correct answer. It can be extended to persona models [20], [21] since different people have their preferences and personalities for dialogue responses.

Ethical Problems

Dialogue systems with emotion regulation can be applied for addressing mental health and counselling situations in the future. However, the text generated from an end-to-end model is a black box, and therefore, the responses generated may be unintended. The problem of unethical dialogue and behavior does not only exist for dialogue systems but for

all humanoid artificial intelligence systems. Methods such as online learning that can perform sequential learning should be considered, including methods personalized for the desired users. Privacy is another issue of concern. A long-term consideration of context and preferences means that one's emotional state is no longer private. Therefore, we expect to be able to run such agents on devices that apply embedded machine learning techniques to protect user privacy while performing context processing entirely locally.

6 Conclusion

This paper proposes ER-Chat, a text-to-text dialogue generation framework for emotion regulation. The proposed method enables dialogue generation by fine-tuning T5 model to predict emotions and intents expressed by listeners based on the context of their dialogues. The proposed method consists of automatic evaluations using Perplexity, BERTScore, and Distinct-N as the evaluation measures for dialogue generation, Likert scale with 100 subjects, and A/B test with 100 subjects based on five evaluation metrics: Fluency, Relevance, Empathy, Emotion Awareness, and Emotion Appropriateness. The results showed that ER-Chat showed comparable performance as that of the SoTA method on the dialogue generation metrics and outperformed the baseline method on human evaluation. By applying the proposed dialogue system to real applications, we believe that it may be possible to realize a dialogue system that is human-like and builds social relationships by storing past dialogues and other information from users over long periods of time.

References

[1] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 3, pp. 1–32, 2020.

[2] R. W. Picard, *Affective computing*, 2000.

[3] H. Prendinger, J. Mori, and M. Ishizuka, "Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game," *International journal of human-computer studies*, vol. 62, no. 2, pp. 231–245, 2005.

[4] H. Prendinger and M. Ishizuka, "The empathic companion: A character-based interface that addresses users'affective states," *Applied Artificial Intelligence*, vol. 19, no. 3-4, pp. 267–285, 2005.

[5] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[6] C. Huang, O. R. Zaiane, A. Trabelsi, and N. Dziri, "Automatic dialogue generation with expressed emotions," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 49–54.

[7] X. Sun, X. Chen, Z. Pei, and F. Ren, "Emotional human machine conversation generation based on seqgan," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018, pp. 1–6.

[8] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: a new benchmark and dataset," Nov. 2018.

[9] Q. Li, H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen, "Empdg: Multi-resolution interactive empathetic dialogue generation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4454–4466.

[10] J. J. Gross, "The emerging field of emotion regulation: An integrative review," *Rev. Gen. Psychol.*, vol. 2, no. 3, pp. 271–299, Sep. 1998.

[11] S. L. Koole, "The psychology of emotion regulation: An integrative review," *Cognition and Emotion*, vol. 23, no. 1, pp. 4–41, Jan. 2009.

[12] C. Reeck, D. R. Ames, and K. N. Ochsner, "The social regulation of emotion: An integrative, Cross-Disciplinary model," *Trends Cogn. Sci.*, vol. 20, no. 1, pp. 47–63, Jan. 2016.

[13] N. M. Thompson, A. Uusberg, J. J. Gross, and B. Chakrabarti, "Empathy and emotion regulation: An integrative account," *Prog. Brain Res.*, vol. 247, pp. 273–304, May 2019.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.

[15] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1577–1586. [Online]. Available: https://aclanthology.org/P15-1152

[16] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," *Advances in NIPS*, 2014.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

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

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[20] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and W. B. Dolan, "A persona-based neural conversation model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 994–1003.

[21] Q. Liu, Y. Chen, B. Chen, J.-G. Lou, Z. Chen, B. Zhou, and D. Zhang, "You impress me: Dialogue generation via mutual persona perception," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1417–1427.

[22] A. W. Li, V. Jiang, S. Y. Feng, J. Sprague, W. Zhou, and J. Hoey, "ALOHA: Artificial learning of human attributes for dialogue agents," *AAAI*, vol. 34, no. 05, pp. 8155–8163, Apr. 2020.

[23] N. Asghar, P. Poupart, J. Hoey, X. Jiang, and L. Mou, "Affective neural response generation," in *European Conference on Information Retrieval*. Springer, 2018, pp. 154–166.

[24] P. Colombo, W. Witon, A. Modi, J. Kennedy, and M. Kapadia, "Affect-driven dialog generation," 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. 3734–3743.

[25] Z. Song, X. Zheng, L. Liu, M. Xu, and X.-J. Huang, "Generating responses with a specific emotion in dialog," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. aclweb.org, 2019, pp. 3685–3695.

[26] "Emotion-regularized conditional variational autoencoder for emotional response generation," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[27] M. Firdaus, H. Chauhan, A. Ekbal, and P. Bhattacharyya, "More the merrier: Towards Multi-Emotion and intensity controllable response generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35. aaai.org, 2021, pp. 12 821–12 829.

[28] F. Cui, H. Di, L. Shen, K. Ouchi, Z. Liu, and J. Xu, "Modeling semantic and emotional relationship in multi-turn emotional conversations using multi-task learning," *Appl. Intell.*, Jul. 2021.

[29] L. Shen and Y. Feng, "CDL: Curriculum dual learning for Emotion-Controllable response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 556–566.

[30] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "MoEL: Mixture of empathetic listeners," Aug. 2019.

[31] N. Majumder, P. Hong, S. Peng, J. Lu, D. Ghosal, A. Gelbukh, R. Mihalcea, and S. Poria, "MIME: MIMicking emotions for empathetic response generation," Oct. 2020.

[32] R. Zandie and M. H. Mahoor, "EmpTransfo: A Multi-Head transformer architecture for creating empathetic dialog systems," in *The Thirty-Third International Flairs Conference*. aaai.org, May 2020.

[33] H.-J. Choi and Y.-J. Lee, "Deep learning based response generation using emotion feature extraction," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2020, pp. 255–262.
[34] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, "Positive emotion elicitation in chat-based dialogue systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 866–877, 2019.
[35] Y. Liang, F. Meng, Y. Zhang, Y. Chen, J. Xu, and J. Zhou, "Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 343–13 352.
[36] Z. Ma, R. Yang, B. Du, and Y. Chen, "A control unit for emotional conversation generation," *IEEE Access*, vol. 8, pp. 43 168–43 176, 2020.
[37] W. Wei, J. Liu, X. Mao, G. Guo, F. Zhu, P. Zhou, and Y. Hu, "Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 1401–1410.
[38] J. J. Gross, H. Uusberg, and A. Uusberg, "Mental illness and well-being: an affect regulation perspective," *World Psychiatry*, vol. 18, no. 2, pp. 130–139, 2019.
[39] K. Niven, P. Totterdell, and D. Holman, "A classification of controlled interpersonal affect regulation strategies." *Emotion*, vol. 9, no. 4, p. 498, 2009.
[40] C. Reeck, D. R. Ames, and K. N. Ochsner, "The social regulation of emotion: An integrative, cross-disciplinary model," *Trends in cognitive sciences*, vol. 20, no. 1, pp. 47–63, 2016.
[41] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.
[42] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.
[43] A. Welivita and P. Pu, "A taxonomy of empathetic response intents in human social conversations," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4886–4899. [Online]. Available: https://aclanthology.org/2020.coling-main.429
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[45] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2019.
[46] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 889–898. [Online]. Available: https://aclanthology.org/P18-1082
[47] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.
[48] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. 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*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2122–2132. [Online]. Available: https://www.aclweb.org/anthology/D16-1230
[49] J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 110–119.
[50] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.
[51] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
[52] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in *Emotion measurement*. Elsevier, 2016, pp. 201–237.

ACKNOWLEDGMENTS

This research was partially supported by JST CREST JPMJCR1882 and NICT.

Shin Katayama is currently a Ph.D. student at Graduate School of Engineering, Nagoya University. He holds a B.A. in Environment and Information Studies (2018) and a M.S. in Media and Governance (2020) from Keio University. His current research interests include human-computer interaction, ubiquitous computing systems, and affective computing.

Shunsuke Aoki received the B.Eng. degree from Waseda University, Tokyo, Japan, the M.S. degree from The University of Tokyo, Tokyo, Japan, and the Ph.D. degree from Carnegie Mellon University, Pittsburgh, PA, USA. After working as a researcher at Carnegie Mellon University and Nagoya University, Japan, he is currently an Assistant Professor at National Institute of Informatics, Tokyo, Japan. His research interests include cyber -physical systems, vehicular communications, and multi-robot coordination.

Takuro Yonezawa is an associate professor in Graduate School of Engineering, Nagoya University, Japan. He received Ph.D. degree in Media and Governance from Keio University in 2010. His research interests are the intersection of the distributed systems, human-computer interaction and sensors/actuators technologies. He is a member of IPSJ, IEICE and ACM.

Tadashi Okoshi is an associate professor in the Faculty of Environment and Information Studies, Keio University. He holds B.A. in Environmental Information (1998) and Master of Media and Governance (2000) from Keio University, M.S. in Computer Science (2006) from Carnegie Mellon University, and Ph.D. in Media and Governance (2015) from Keio University, respectively. His recent research interests are human attention-awareness and management in ubiquitous computing and cyber physical systems. In Keio University and Carnegie Mellon University, he has been working on mobile and ubiquitous computing systems, actively joining several RD projects.

Jin Nakazawa is professor at Faculty of Environment and Information Studies at Keio University, Japan. He has received his Bachelor's degree (1998) in Faculty of Policy Management, Master's degree (2000), and Doctor of Philosophy degree (2003) in Media and Governance from Keio University. His research interest includes middleware systems, distributed systems, ubiquitous computing systems, and life-logging. He is a member of IEEE, IEICE, and IPSJ.

Nobuo Kawaguchi received his B.E., M.E. and Ph.D. degrees in computer science from Nagoya University, Japan, in 1990, 1992, and 1995, respectively. From 1995, he was an associate professor in the Department of Electrical and Electronic Engineering and Information Engineering, School of Engineering, Nagoya University. Since 2009, he has been a professor in the graduate school of Engineering, Nagoya University. His research interests are in the areas of recognition of human activity, smart environmental systems, and ubiquitous communication systems.