IEICE TRANS. INF. & SYST., VOL.E105–D, NO.5 MAY 2022

PAPER Special Section on Data Engineering and Information Management

Toward Generating Robot-Robot Natural Counseling Dialogue

Tomoya HASHIGUCHI[†a)], ***Nonmember***, **Takehiro YAMAMOTO**[††,†††], ***Member***, **Sumio FUJITA**[††††], ***Nonmember***, ***and*** **Hiroaki OHSHIMA**[†,††,†††], ***Member***

SUMMARY In this study, we generate dialogue contents in which two systems discuss their distress with each other. The user inputs sentences that include environment and feelings of distress. The system generates the dialogue content from the input. In this study, we created dialogue data about distress in order to generate them using deep learning. The generative model fine-tunes the GPT of the pre-trained model using the TransferTransfo method. The contribution of this study is the creation of a conversational dataset using publicly available data. This study used EmpatheticDialogues, an existing empathetic dialogue dataset, and Reddit r/offmychest, a public data set of distress. The models fine-tuned with each data were evaluated both automatically (such as by the BLEU and ROUGE scores) and manually (such as by relevance and empathy) by human assessors.

key words: *dialogue system, robot-robot interaction*

1. Introduction

In this study, we tackle the problem of generating the dialogue content in which one robot having distress consult another robot. The example of the dialogue to be created in this study is shown in Fig. 1. The system first receives the situation that describes the environment and emotions of the user's distress. Then the system generates a dialogue between two robots. For example, the input situation is "I want to quit my job because my relationships at work have not been good lately." The speaking robot (we refer to the robot as Speaker hereafter) first utters, "I want to quit my current job." The listening robot (we refer to the robot as Listener hereafter) then responds as "Year, I know it's a lot to take in. Why do you want to quit." The Speaker then responds, "I don't feel comfortable at work because ..." In our system, the user passively watches the generated robot-robot dialogue. The proposed robot-robot interaction is not unique. Hayashi et al. proposed three types of interactions between the system and the user [1], [2], as shown in Fig. 2. They hypothesize that it is more natural and understandable to the robot-to-robot system (social-passive) than a single system (passive).

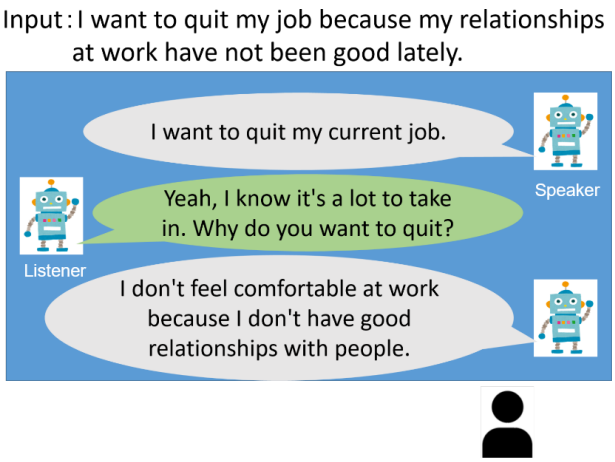

Fig. 1 Image of the dialogue content.

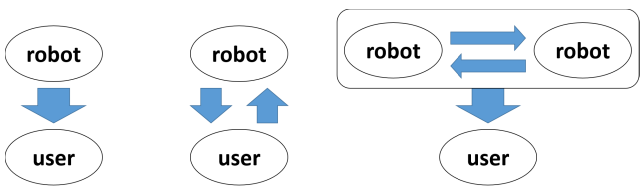

Fig. 2 Human-System Interaction (Figure created based on Hayashi et al. [1], [2]).

There are two reasons why we used this interaction. The first reason is to avoid creating user-driven content. People's distress and answers on Yahoo! Answers and Reddit are a kind of content. A large number of users see content, a few users comment or post them. For this reason, we thought that users would enjoy browsing content rather than taking the initiative to create content as in a dialogue system. The second reason is that the system has a high hurdle for discussing problems. There is a great deal of research on dialogue systems that empathize with the user's utterance. Some of the studies also provide counseling for depressed patients' problems. The cooperation of psychologists and medical experts is essential for such a system to help users with their specific concerns. In this research, we expect to create a peer support effect in which users perceive the system's conversations as tweets and feel that other people are in a similar situation.

The recent development in NLP such as GPT and GPT-2 enables us to generate natural dialogue [3], [4]. However, when generating a dialogue about distress, the avail-

Manuscript received June 26, 2021.
Manuscript revised November 16, 2021.
Manuscript publicized February 7, 2022.

[†]The authors are with Graduate School of Applied Informatics, University of Hyogo, Kobe-shi, Hyogo, 650–0047 Japan.
[††]The authors are with School of Social Information Science, University of Hyogo, Kobe-shi, Hyogo, 651–2197 Japan.
[†††]The authors are with Graduate School of Information Science, University of Hyogo, Kobe-shi, Hyogo, 650–0047 Japan.
[††††]The author is with Yahoo Japan Corporation, Tokyo, 102–8282 Japan.
a) E-mail: thashiguchi1995@gmail.com

DOI: 10.1587/transinf.2021DAP0008

able dataset for training the model is limited and tiny. One of the datasets that can be used to train the model in our study is EmpatheticDialogues [5]. EmpatheticDialogues is a crowdsourced dataset to evaluate whether a dialogue system shows empathetic responses. However, the number of conversations about distress in the dataset is very few. Furthermore, creating such a dataset by crowdsourcing requires some cost in terms of time and money.

Therefore, in this study, we propose to create the dialog data for training the model from Reddit. Our proposed method creates the dialogue by using the response structure in Reddit. We compare the EmpatheticDialogues dataset and the dataset extracted from the Reddit for generating the dialogue about distress consultation.

The followings are the contributions of this study:

- We proposed a method for automatically extracting the dialogue data from Reddit by using its response structures. Our method would require less effort in preparing the training data for training the model compared with the existing approach that uses crowdsourcing. Also, from the experiments, we showed that our method that uses the data extracted from Reddit can generate dialogues as the same as the baseline method that uses the EmpatheticDialogues dataset in terms of BLEU, ROUGE, length of turns, and human evaluation.
- We proposed the idea of generating a dialogue about worry between two robots in which one robot having a worry consults another robot.

2. Related Work

2.1 Empathy/Sympathy

In this study, we focus on people who are having problems. Dialogue generation for this user often shows empathy and sympathy by the dialogue system. One of the datasets used in this study is the dataset for assessing empathy built by Rashkin et al. [5]. Rashkin et al. collected dialogue data through crowdsourcing. Specifically, one of the two crowd workers assumes a particular emotional situation and starts a conversation. Another crowd worker responds to the content, building an empathic dialogue data set. Zhong et al. believe that in addition to emotion, user persona is also important for empathy [6]. For this reason, Zhong et al. created PEC (Persona-based Empathetic Conversation) as a conversation dataset with emotion and persona. In addition, understanding emotional responses and showing compassion in chitchat is said to improve performance on many tasks [7], [8].

Kim et al. are working on issues to reduce bias against depressed patients. [9]. Kim et al. create a virtual character to impersonate a depressed person. In addition, Kim et al. showed that interacting with a virtual character can help people empathize with the character's worries and reconstruct their own worries and problems, increase their motivation to help others in need, and reduce prejudice against depression. On the other hand, we also consider the burden of negative topics on the supporters as an issue. In this study, we propose a dialogue between a character who has a problem and a character who listens to the character's problem, so that the user can talk about his or her problem without being overwhelmed.

2.2 User and System Interaction

In this research, we provide users with a conversation robot-to-robot. There are several studies using such system-user interaction. As mentioned in Sect. 1, Hayashi et al. define the following three types of interactions between a user and a system.

- Interactions that receive information unilaterally from the system
- Interactions that interact with the system and each other
- Interaction to observe how systems interact with each other

Hayashi et al. created a robot that performs boke (comedy) and tsukomi (comedy) to allow users to observe comic performances [1], [2].

2.3 Generation Method

In this research, we will use language generation techniques to create content. In the past, RNNs and CNNs were commonly used for language generation [10]–[14]. Vaswani et al. proposed Transformer as a model that uses only the Attention mechanism for generation, not RNN or CNN, and achieved the highest accuracy [15]. Radford et al. proposed GPT (Generative Pre-Training) as a generative model using Transformer's model [16]. Wolf et al. fine-tuned the GPT model for dialogue generation and achieved the highest accuracy in ConvAI2, a dialogue competition held at NeurIPS2018 [3]. In this study, we use the TransferTransfo learning method. Zhang et al. also proposed DialoGPT as a model fine-tuned for dialogue generation in GPT2, an improved version of GPT [4].

3. Dialogue Generation Method

The input-output format in this study is shown in Fig. 3. As mentioned in Sect. 1, The input is a sentence that includes the environment and feelings of distress that the user is having. In order to generate conversations for each SPEAKER

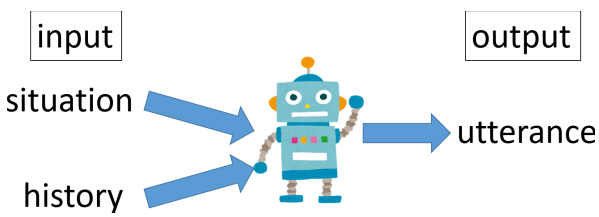

Fig. 3 Input and output for dialogue content generation.

and LISTENER, the history of the generated conversations is also input. In this study, dialogue contents are generated by sequencing the output utterances.

3.1 GPT

GPT (Generative Pre-Training) proposed by Radford et al. is a pre-training model with multiple layers of Decoder layers in the Transformer [16]. GPT is a model used in language generation. The GPT pre-training is done by learning to predict the next word using BooksCorpus, which contains over 7000 books in various genres such as adventure, fantasy, and romance.

GPT can also adapt the model to the task by fine-tuning it as in BERT [17]with a layer of Transformer Encoder. Fine-tuning with GPT is a multi-tasking task that consists of a language model task to predict the next word and an intrinsic task to be fine-tuned. The calculation of the loss value of multitasking when fine-tuning a GPT is as follows:

$$\text{loss} = rate_{lm} \cdot loss_{lm} + rate_{task} \cdot loss_{task} \quad (1)$$

In this study, we fine-tune the GPT by using $loss_{lm}$ as the loss value in the task of generating the utterance as in the TransferTransfo method, and $loss_{task}$ as the loss value in the task of determining whether the utterance is contextually correct. In addition, $rate_{lm}$ and $rate_{task}$ were set as hyperparameters, respectively.

3.2 GPT Fine-Tuning

TransferTransfo proposed by Wolf et al. is the model that achieved the highest accuracy in ConvAI2, a dialogue competition held at NeurIPS2018 [3]. In this study, we use TransferTransfo to input the situation in Fig. 4 and generate the dialogue between SPEAKER and LISTENER. The fine-tuning of the GPT in this study is done as shown in Fig. 5. The model inputs the situation and conversation history into the GPT and learns a language model task that predicts the next word to be uttered and an intrinsic task that

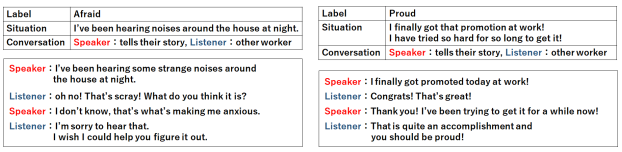

Fig. 4 EmpatheticDialogues dataset. (Figure based on Rashkin et al [5])

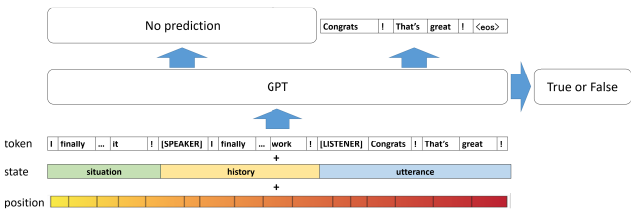

Fig. 5 Learning to generate dialogue content for distress.

classifies whether the utterance is appropriate. The history of a conversation consists of a pair of the utterances between SPEAKER and LISTENER.

4. Dataset

We use two datasets for fine-tuning GPT in the input-output format of TransferTransfo. We used EmpatheticDialogues as an existing dialogue dataset for empathy, and posts and comments in the Reddit r/offmychest as a dataset where the distress are discussed.

4.1 EmpatheticDialogues

EmpatheticDialogues is a dataset created for the evaluation of dialogue systems that empathize with the user's statements. EmpatheticDialogues uses crowdsourcing to create data from a dialogue between two workers. One worker selects one of 32 emotional items such as surprise, anger, or fun, and the other worker decides on a topic that matches the item. For example, for surprise, a worker can say, "My friend surprised me on my birthday," and for anger, a worker can say, "My friend broke an appointment that I was looking forward to." The task of the worker is to interact with the other worker on the topic that has been decided. The datasets are also divided into training, validation, and testing. In this study, we use this dataset in a problem setting where the user inputs an emotional situation and generates a conversation based on the input situation.

4.2 Reddit r/offmychest

The EmpatheticDialogues dataset has been collected through crowdsourcing. Therefore, it may not be appropriate for the listener to respond. In this study, we also used data from Reddit as a user-contributed community site. Here, the contributor is the user who created the thread, the post is the text of the contributor's thread, and the comment is the text sent to the post.

Reddit r/offmychest is where users post their problems that they find difficult to talk about with their acquaintances. In addition, other users who are interested in such posts make comments on them, and the contributors make comments on those comments in a multi-turn exchange. In this study, we use the data of posts and comments shown in red boxes in Fig. 6 as multi-turn data.

The r/offmychest has been widely used in existing research to empathize with users' negative emotions.

In this study, we use the Reddit r/offmychest data [†] collected by Jaidka et al. [18]. Data collected in existing studies were not commented on for submission, so additional data were collected. The situation given as input was taken from the first post of the trouble contributor. In order to match the EmpatheticDialogues data as closely as possible, the sentences of the situation were set to three. In the case of three

[†]https://github.com/kj2013/claff-offmychest

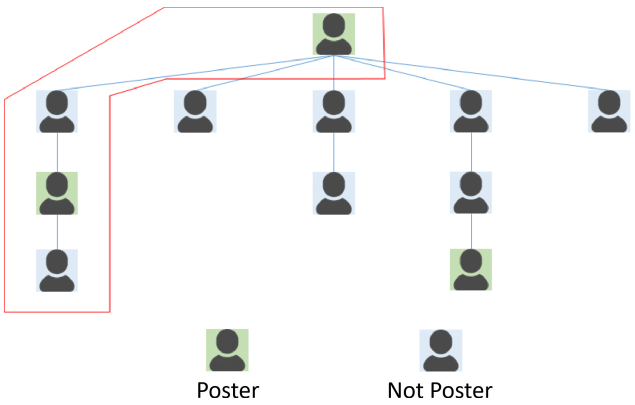

Fig. 6 Conversation data from Reddit r/offmychest.

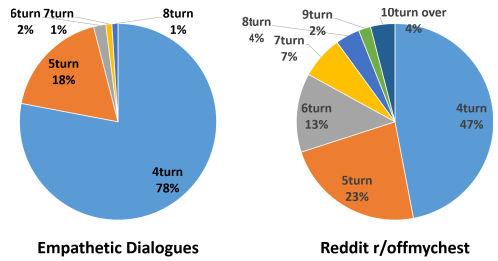

Fig. 7 Percentage of conversation data by number of turns.

Fig. 8 Input without conversation history.

Fig. 9 Input with conversation history.

sentences or less, the entire submission was considered a situation statement. In this study, posts were divided into sentence units using spaCy [†].

4.3 Number of Utterances in a Conversation

The percentage of multi-turn conversations in each dataset is shown in Fig. 7. The total number of multi-turns in each dataset is 19,532 for EmpatheticDialogues and 6,636 for Reddit.

4.4 Data Creation for Content Generation

The training data of the model is in a similar input/output format to use TransferTransfo. The input/output format of TransferTransfo in this research is the user's situation and conversation history as input, and the next utterance as output.

A part of the training data of TransferTransfo is shown in Fig. 8 and Fig. 9. The data created will be as follows.

$$\text{convs} = (situation_1, conv_1) \cdots (situation_n, conv_n)$$

Here, we have $situation_i = \{s_1 \cdots s_n\}$ is the user's situation given as input, and where $s_1, \cdots, s_n$ are the multiple sentences in $situation_i$.

Also.

$$conv_i = (None, u_1), (u_1, u_2) \cdots (u_{1,n-1}, u_n)$$

is the conversation that took place in $situation_i$, u_n is the utterance that took place in the nth turn, and $u_{1,n-1}$ is the history of conversations that took place before the nth turn.

Since the first utterance has no history, it was set to None. In addition, the utterance is as follows

$$u_i = (u_{false}, \cdots u_{false}, u_{true})$$

The u_{false} is a randomly obtained negative sample utterance, and u_{true} is the correct answer utterance.

The training data was created using the following procedure. First, divide the data into conversational units, and create pairs of situations and conversations. The situation and conversation pairs are then divided 8:1:1 for training, validation, and testing. Next, the conversation is divided into turn-by-turn, conversation history and correct utterances, and negative examples are created by extracting random utterances in the data.

The generation of the dialogue stops (1) when the Speaker's generated utterance is the same as the Speaker's own utterances in the past, or (2) when the Listener's generated utterance is the same as the Listener's own utterances in the past. The utterances just before the same utterance is generated are uses as the output of the dialogue. Let S_i be the Speaker's ith uttrance and L_i be the Listener's ith uttrance. For example, if the generate uttrances are S_1 = "hello", L_1 = "what's up", S_2 = "I have a headache", L_2 = "that's too bad", S_3 = "I have a headache" then the utterances up to S_1, L_1, S_2, L_2 are the output as the dialogue.

4.5 Examples of Generated Dialogue

The example generated dialogue is shown in Fig. 10. The dialogue was generated based on the input situation selected from the question having a distress in workplace posted in Yahoo! Chiebukuro[††]. We translated the question into English and used it as the input situation of the model. From the figure, we can see that the model trained on the Reddit r/offmychest dataset generated a longer dialogue compared with the one on the EmphatheticDialogues.

From the figure, we can see that the model trained on the Reddit r/offmychest dataset generated a longer dialogue compared with the one on the Emphathetic Dialogues.

[†]https://spacy.io/

[††]https://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q11225464071

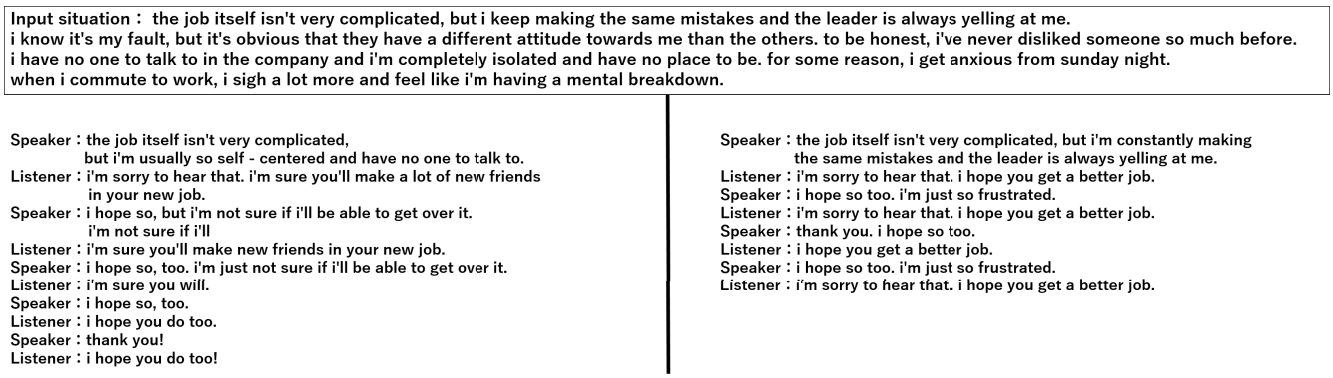

Fig. 10 Generated dialogue about a distress in workplace. Left: Trained on Reddit r/offmychest dataset, right: trained on the EmpatheticDialogues dataset.

5. Experiments

To evaluate the quality of the generated dialogue, we employed both the automatic and manual evaluation. We first explains the experimental setting and then reports the results.

5.1 Methods and Implementation

In this study, we compared the two models. A baseline model uses the TransferTransfo model trained on the EmpatheticDialogues dataset. We refer to the model as "EmpatheticDialogues". Our proposed model uses the TransferTransfo model trained on the Reddit r/offmychest dataset created in this study. We refer to the model as "Reddit".

The hyperparameters used for training for both models are shown in Table 1. Other hyperparameters used were the default values of the TransferTransfo implementation [†]. We used PyTorch (version 1.3) to implement the TransferTransfo model. We used four Tesla V100 GPUs (16GB) for the training.

The hyperparameters for generating the dialogue such as *top_k*, *top_p*, *temperature*, *max_history* were the same as the TransferTransfo model. The test data of the EmpatheticDialogues dataset was used to evaluate the models. The dialogue generation continued until either of the SPEAKER or LISTENER generated the same utterance as the previous one.

5.2 Automatic Evaluation

As for the automatic evaluation, BLEU and ROUGE scores are used to evaluate whether the generated dialogue matches the ground truth dialogue [††]. Our proposed method prepares the dialogues for training from the structure of Reddit responses. This means that our method implicitly generates the dialogue content while the baseline method explicitly prepares the dialogues by using the crowd workers. Since both BLEU and ROUGE become higher when the words (n-grams) in the generated text match the ground truth, we could use BLEU and ROUGE to measure how close the generated utterances are to human response to some extent. BLEU and ROUGE have been commonly used in dialogue evaluation [19]–[21]. One of the limitations of BLEU and ROUGE is they rely on simple word matching. There may be cases where the generated dialogue is natural even if the words do not match the ground truth. So, we will also conduct a manual evaluation as in Sect. 5.3.

Table 1 Hyperparameters used for fine-tuning

Hyperparameters	Values
max. number of input history	2
Batch size	2
learning rate	6.25×10^{-5}
epoch	3
rate_lm	2.0
rate_task	1.0

The results of the evaluation of BLEU and ROUGE scores using the situation of the EmpatheticDialogues test data are shown in Table 2. From the figure, we can see that the BLEU and ROUGE scores of the baseline and proposed models are comparable. These results indicate that our method, which is trained on the dialogue implicitly generated from Reddit, can generate dialogues as effective as the baseline method in terms of BLEU and ROUGE. Note that the BLEU and ROUGE were evaluated using the test data of the EmpatheticDialogues dataset, the baseline method, which is trained on the training data in the EmpatheticDialogues dataset, is more likely to obtain higher BLEU and ROUGE scores compared with those obtained by the proposed method. This suggests that a dialogue corpus generated from Reddit can be used as training data for learning the model.

Table 3 shows the length of turns of the generated dialogues. Also, we compute the length of turns until the dialogue ends. The speaker/listener's length of turns is defined as the number of utterances generated by the speaker/listener. For example, in the example described in Sect. 4.4, the speaker's length of turns is 2, and the listener's length of turns is also 2. The more the length of turns, the longer the conversation lasts, and the less the length of turns, the shorter the conversation ends. In this study, since the content was created through the dialogue between

[†]https://github.com/huggingface/transfer-learning-conv-ai
[††]https://github.com/chakki-works/sumeval

Table 2 BLEU and ROUGE scores of generated dialogue

	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-BE
EmpatheticDialogues	5.84	0.177	0.086	0.168	0.088
Reddit	5.97	0.180	0.088	0.171	0.089

Table 3 Turn length of the generated dialogue

	Turn length	
	Speaker	Listener
EmpatheticDialogues	4.23	3.50
Reddit	4.05	3.30

Table 4 Model(Reddit) results from human evaluation

	score1	score2	score3	score4	score5
Relevance(LISTENER)	31(6.2%)	54(10.8%)	140(28.0%)	190(38.0%)	85(17.0%)
Relevance(SPEAKER)	30(6.0%)	63(12.6%)	79(15.8%)	184(36.8%)	144(28.8%)
Empathy / Sympathy	27(5.4%)	41(8.2%)	117(23.4%)	205(41.0%)	110(22.0%)
Fluency	28(5.6%)	53(10.6%)	114(22.8%)	151(30.2%)	154(30.8%)

Table 5 Model(EmpatheticDialogues) results from human evaluation

	score1	score2	score3	score4	score5
Relevance(LISTENER)	17(3.4%)	59(11.8%)	112(22.4%)	225(45.0%)	86(17.2%)
Relevance(SPEAKER)	18(3.6%)	53(10.6%)	104(20.8%)	163(32.6%)	161(32.2%)
Empathy / Sympathy	14(2.8%)	48(9.6%)	124(24.8%)	182(36.4%)	131(26.2%)
Fluency	18(3.6%)	41(8.2%)	123(24.6%)	177(35.4%)	140(28.0%)

Table 6 Mean and standard deviation of the human evaluation

	Model(Reddit)		Model(ED)		
	Mean	SD	Mean	SD	p-value
Relevance(LISTENER)	3.48	1.18	3.61	1.02	0.07
Relevance(SPEAKER)	3.69	1.40	3.79	1.23	0.18
Empathy / Sympathy	3.66	1.15	3.74	1.08	0.23
Fluency	3.70	1.37	3.76	1.12	0.39

robots, we consider the generated dialogue ineffective when its length of turns is short, such as once or twice. The results in Table 3 show that the proposed method using Reddit has the same length of turns as the baseline using Empathetic Dialogues, and both methods have more than four turns for the speaker and more than three turns for the listener. Therefore, we can say that, in terms of the length of turns, our method, which trains the model by the data automatically constructed from the Reddit posts, is as effective as the EmpatheticDialogues dataset, in which the dialogue data is manually created by the crowdsourcing.

5.3 Manual Evaluation

As for the manual evaluation, we evaluated the relevance, empathy/sympathy, and fluency of the generated dialogue. We used Amazon Mechanical Turk for human evaluation of the generated dialogue. The followings are the instructions used in the experiment.

Task Description: The following is a conversation between a person who has a problem (SPEAKER) and a person who is receiving the advice (LISTENER). You are to rate this conversation on a scale of 1 to 5 for each of the four evaluation items. There are four evaluation items as follows

- Relevance(LISTENER):Did the responses of the LISTENER seem appropriate to the conversation? Were they on topic?
- Relevance(SPEAKER):Did the responses of the SPEAKER seem appropriate to the conversation? Were they on topic?
- Empathy / Sympathy: Did the responses from the LISTENER show understanding of the feelings of the SPEAKER talking about distress?
- Fluency: Could you understand the dialog content? Did the language seem accurate?

The number of assessors per task was five. The input situations for both models are selected as follows. As for the EmpatheticDialogues model, the input situations were sampled from the test data in the EmpatheticDialogues dataset. As for the Reddit model, we collected the posts posted in Reddit between March 12 and March 15, 2021. In order to avoid that the input subject is not related to the distress, we only used the negative situation, which is identified by the sentiment-analysis of the model provided by Huggingface[†]. We generated 50 dialogues for both models and used them in the human evaluation.

We asked the assessors to evaluate each of the following categories from 1 to 5. 5 is the best score.

- Relevance(LISTENER)
- Relevance(SPEAKER)
- Empathy / Sympathy
- Fluency

In total, we obtained 498 evaluations. The manual evaluation results are shown in Table 4 and Table 5.

[†]https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

Tables 4 and 5 shows the distributions of the scores for two methods. Table 6 shows the mean and standard deviation for each method. The results of Welch's t-test was also shown in the table. The results of Table 6 showed no significant difference between the proposed method and the comparison method. However, the baseline method, which uses the EmpatheticDialogues dataset, requires manual effort in creating the dialogue data by crowdsourcing. In contrast, the proposed method can create the dialogue data from Reddit automatically. Our method achieved a similar quality in terms of the naturalness of the dialog, while our method would require less effort in creating the dataset for training the model.

This result was different from the hypothesis of this study, that Reddit is better at generating Empatheticdialogues. As a factor for this, we thought that EmpatheticDialogues may be able to generate better because the input sentences of EmpatheticDialogues are simplified by summarizing what people want to talk about. One possible improvement would be to change the input method when using Reddit. In this study, the first three sentences were used, but it is possible to summarize them and use them as input sentences.

6. Conclusion and Discussion

The goal of this study was to generate a dialogue about worry between two robots in which one robot having a worry consults another robot. To achieve this goal, we proposed a method to extract the dialogue data from Reddit automatically. We found that the effectiveness of our method was comparable to that using the Empathetic Dialogues dataset in terms of BLEU, ROUGE, the length of the dialogue, and human evaluation. One way to generate a more natural dialogue of worries would be to prepare the different models for speakers and listeners. In this study, the speaker and listener are trained with a single model.

Another important research direction would be evaluating the usefulness of our proposed idea of generating a dialogue about worry between two robots in which one robot having a worry consults another robot. The current study only evaluated the quality of the generated dialogue. We have not evaluated how much people prefer the robot-robot dialogue about worry. We need to examine how our idea of robot-robot natural counseling dialogue has benefit for people having a distress.

Acknowledgments

The work was supported by the Research Grant by JSPS KAKENHI Grant Numbers JP18H03494, JP21H03774, JP21H03775, JP21H03554.

References

[1] K. Hayashi, D. Sakamoto, T. Kanda, M. Shiomi, S. Koizumi, H. Ishiguro, T. Ogasawara, and N. Hagita, "Humanoid robots as a passive-social medium: a field experiment at a train station," Proc. HRI'07, pp.137–144, March 2007.

[2] K. Hayashi, T. Kanda, T. Miyashita, H. Ishiguro, and N. Hagita, "Robot manzai: Robot conversation as a passive-social medium," Int. J. Humanoid Robotics, vol.5, no.1, pp.67–86, 2008.

[3] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "TransferTransfo: A Transfer learning approach for neural network based conversational agents," Proc. NIPS'18, pp.1–6, 2018.

[4] Y. Zhang, S. Sun, M. Galley, Y.C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DialoGPT: Large-scale generative pre-training for conversational response generation," Proc. ACL'2020 system demonstration, pp.270–279, 2020.

[5] H. Rashkin, E.M. Smith, M. Li, and Y.L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," Proc. ACL'19, pp.5370–5381, 2019.

[6] P. Zhong, Y. Zhu, Y. Liu, C. Zhang, H. Wang, Z. Nie, and C. Miao, "Towards persona-based empathetic conversational models," Proc. EMNLP'20, pp.6556–6566, 2020.

[7] W. Levinson, R. Gorawara-Bhat, and J. Lamb, "A Study of patient clues and physician responses in primary care and surgical settings," JAMA, vol.284, no.8, pp.1021–1027, 2000.

[8] K. Wentzel, "Student motivation in middle school: The role of perceived pedagogical caring.," J. Educational Psychology, vol.89, no.3, pp.411–419, 1997.

[9] T. Kim, M. Ruensuk, and H. Hong, "In helping a vulnerable bot, you help yourself: Designing a social bot as a care-receiver to promote mental health and reduce stigma," Proc. CHI'20, pp.1–13, April 2020.

[10] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," Proc. EMNLP'14, pp.1724–1734, Oct. 2014.

[11] I. Sutskever, O. Vinyals, and Q.V. Le, "Sequence to sequence learning with neural networks," Proc. NIPS'14, pp.3104–3112, Dec. 2014.

[12] O. Vinyals and Q. Le, "A neural conversational model," Proc. ICML'15 Deep Learning Workshop, pp.1–7, 2015.

[13] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.Y. Nie, "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion," Proc. CIKM'15, pp.553–562, Oct. 2015.

[14] I.V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," Proc. AAAI'17, pp.3295–3301, Feb. 2017.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," Proc. NIPS'17, pp.6000–6010, 2017.

[16] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," pp.1–12, 2018.

[17] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL'19, pp.4171–4186, 2019.

[18] K. Jaidka, I. Singh, J. Lu, N. Chhaya, and L. Ungar, "A report of the CL-Aff OffMyChest Shared Task: Modeling supportiveness and disclosure," Proc. AAAI'20 Workshop on Affective Content Analysis, pp.1–12, 2020.

[19] Q. Li, H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen, "EmpDG: Multi-resolution interactive empathetic dialogue generation," Proc. COLING'2020, pp.4454–4466, Dec. 2020.

[20] S. Roller, Y.L. Boureau, J. Weston, A. Bordes, E. Dinan, A. Fan, D. Gunning, D. Ju, M. Li, S. Poff, P. Ringshia, K. Shuster, E.M. Smith, A.D. Szlam, J. Urbanek, and M. Williamson, "Open-domain conversational agents: Current progress, open problems, and future directions," arXiv:2006.12442, 2020.

[21] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre,

and M. Cieliebak, “Survey on evaluation methods for dialogue systems,” Artificial Intelligence Review, vol.54, pp.755–810, 2020.

Tomoya Hashiguchi received his B.S. degree from Department of Intelligence and Informatics, Konan University, in 2019 and M.S. degree from Graduate School of Applied Informatics, University of Hyogo, in 2021. His research interests include dialog systems.

Takehiro Yamamoto received his B.S., M.S., and Ph.D. degrees from Kyoto University, in 2007, 2008 and 2011, respectively. He joined University of Hyogo in 2019, and he currently works as an associate professor of Graduate School of Information Science, University of Hyogo. His research interests include Information Retrieval, Human-Computer Interaction, and Data Mining. He is a member of the Information Processing Society of Japan (IPSJ), and the Database Society of Japan (DBSJ).

Sumio Fujita worked for Computer Institute of Japan Ltd. 1985–1995, studied for DEA at Université de Paris 7 1988–1989, worked as research associate at University of Manchester Institute of Science and Technology 1993–1994, worked as chief researcher at Justsystem Corporation 1995–2002, worked as research scientist at Claritech Corporation 1998, and worked as senior research scientist at Patolis Corporation 2002–2004. He joined Yahoo Japan Corporation in 2005, participated in the foundation of Yahoo! JAPAN Research in 2007. He currently works as project researcher on information retrieval, web mining and related areas at Yahoo! JAPAN Research. He is a member of ACM, SIGIR and IPSJ.

Hiroaki Ohshima received the BS and MS degrees in Engineering from Kobe University, in 2000 and 2003, respectively. In 2006, he received the Ph.D. degree in Informatics from Kyoto University. He worked for the Department of Social Informatics, Graduate School of Informatics, Kyoto University from 2006 to 2017, as an associate professor and a program-specific associate professor. He joined University of Hyogo in 2017, and he currently works as an associate professor in Graduate School of Information Science, University of Hyogo. His research interests include information retrieval, Web search and mining, and information design. He is a member of the ACM, the Information Processing Society of Japan (IPSJ), and the Database Society of Japan (DBSJ).