Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 491–497
Marseille, 11–16 May 2020

Emotional Speech Corpus for Persuasive Dialogue System

Sara Asai[1], Koichiro Yoshino[1,2,3], Seitaro Shinagawa[1,3], Sakriani Sakti[1,3], Satoshi Nakamura[1,3]
[1]Nara Institute of Science and Technology, Takayama 8916-5, Ikoma, Nara, 6300192, Japan
[2]PRESTO, Japan Science and Technology Agency, Saitama, Japan
[3]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan
{koichiro, shinagawa.seitaro.si8, ssakti, s-nakamura} at is.naist.jp

Abstract

Expressing emotion is known as an efficient way to persuade one's dialogue partner to accept one's claim or proposal. Emotional expression in speech can express the speaker's emotion more directly than using only emotion expression in text, which will lead to a more persuasive dialogue. In this paper, we built a speech dialogue corpus in a persuasive scenario that uses emotional expressions to build a persuasive dialogue system with emotional expressions. We extended an existing text dialogue corpus by adding variations of emotional responses to cover different combinations of broad dialogue context and a variety of emotional states by crowd-sourcing. Then, we recorded emotional speech consisting of collected emotional expressions spoken by a voice actor. The experimental results indicate that the collected emotional expressions with their speeches have higher emotional expressiveness for expressing the system's emotion to users.

Keywords: dialogue corpus, emotional expression, speech corpus, persuasive dialogue

1. Introduction

Persuasion or negotiation is an important dialogue style, which has been widely researched recently (Mazzotta et al., 2007; Georgila, 2013; Hiraoka et al., 2016; Wang et al., 2019). Emotional expressions have an important role in various dialogue situations and contexts (Keltner and Haidt, 1999; Morris and Keltner, 2000; Adler et al., 2016). It is well known that emotional expressions are useful in persuasion and negotiation (Fogg, 1999): building a cooperative relationship with positive expressions (Forgas, 1998) or pressing the dialogue partner to accept a proposal with negative expressions (Sinaceur and Tiedens, 2006). We built dialogue corpora in a persuasive scenario annotated with emotion labels to build persuasive dialogue systems that can use emotional expressions to improve its success rate (Yoshino et al., 2018).

When the system uses emotional expression in a dialogue, it is important to correctly express the emotion that the system intended. Expressing actual emotion to the users with only textual information is sometimes difficult because textual information has limited expressiveness. In contrast, emotional speech or gesture has the potential to improve the expressiveness for expressing the intended emotional state to the user.

In this paper, we collected emotional expressions for the persuasive scenario and recorded their audio by expressing the emotional state to be indicated to the dialogue partner. Existing dialogue corpora based on natural scenario collection (Yoshino et al., 2018) do not have comprehensive emotional expressions for any dialogue contexts. However, for building a dialogue system that can use any emotional states in any dialogue contexts, we collected dialogue responses based on any emotional states given a dialogue context via crowd-sourcing. We converted the sentences in the dialogue contexts into a vector in latent space by using bidirectional encoder representations from Transformers (BERT) (Devlin et al., 2019) for building a robust dialogue corpus. The dialogue contexts to be used for the annotation are selected by K-means clustering to cover possible dialogue contexts.

Emotion	Populations
Neutral	33.54%
Angry	22.72%
Sad	20.91%
Happy	10.57%
Content	3.15%
None	9.11%

Table 1: Proportion of each emotion label of system utterances in existing persuasive dialogue corpora with emotional expressions. Labels are given by agreement of three annotator; thus, "None" indicates samples that was not agreed in the annotation process.

We recorded emotional speech to collect the emotional responses spoken by a well-trained voice actor. We showed our dialogue platform robot "CommU" with its gestures corresponding to emotion classes for making emotional expressions in speech. We also showed Russell ' s circumplex model (Russell, 1978) for indicating the emotion to be expressed in the recording. We evaluated the collected texts and speeches from the viewpoint of the expressiveness of the given emotion.

2. Collection of Emotional Sentences via Crowd-Sourcing

In our previous work to build persuasive dialogue corpora with emotional expressions, the populations of emotion labels in the corpora are biased (Yoshino et al., 2018). Table 1 shows proportion of annotated emotion labels in this paper. Such bias makes it difficult to generate or select a natural response to some given pairs of a dialogue context and an emotion label. To solve this problem, in this work, we extended the existing persuasive dialogue corpora using emotional expressions with additional response variations given different emotion labels.

When we build a dialogue system based on statistical approaches, the system selects or generates a response given a dialogue context. In the selective approach, the system selects response r^t given context q^t from example candidates $< q_i, r_i >$, in time-step t. Here, r_i is the corresponding response to context q_i.

$$\hat{q_i} = \underset{i}{\mathrm{argmax}}(\mathrm{Sim}(q_i, q^t)). \tag{1}$$

Once $\hat{q_i}$, which is the example that is the most similar to the current context q^t, is selected by used similarity function (e.g., cosine-similarity), corresponding response $\hat{r_i}$ is used as response r^t. In the generative approach (Ghosh et al., 2017; Zhou et al., 2018), the system tries to learn a function,

$$r_i = f(q_i), \tag{2}$$

from given training data to find $r^t = f(q^t)$. Each method requires a large-scale corpus $Q(q_i \in Q)$ that is large enough to cover potential dialogue contexts q^t.

In our scenario, the dialogue system has an emotional state and it affects the selection or generation result. In other words, emotional state e^t is used as a given condition in addition to general dialogue context q^t, which increases the data sparsity problem. In dialogue corpora collected as natural conversations, it is difficult to cover any variations of emotional states for the given dialogue context. To prevent this problem, we collected paraphrases of target system responses by giving emotion labels that were different from the emotion label in the original corpora. We show an example in Table 2.

In Table 2, "dialogue context" and "target response" are utterances contained in the original corpus, and "response variations in different emotions" are a new part we collected in this work. The original dialogue corpora have emotion label annotations (five classes: neutral, angry, sad, happy, content) given by three annotators. In this work, we collected response variations for emotion labels that are not annotated to the original "target response" with paraphrasing. Note that there are fewer emotion labels for "content" than other emotions as shown in Table 1 (3.15%), and it is difficult to distinguish them from the "neutral" label; thus, we unified the "content" class label and the "neutral" class label.

The original corpora consist of five different domains of persuasive dialogue: clean the room (cleaning), do not leave a dish unfinished (lunch), sleep early (sleep), stop playing the game (game), and get some exercise (exercise). In this work, we only extended the "exercise" scenario because using a recording audio for any scenario is costly. The original emotion labels are annotated by three annotators; thus, we only used samples that have the same emotion label from two or three annotators. We removed other samples from the annotation. We also removed utterances consisting of silence symbols (...) [1]. The number of resultant utterances that are classified as "target responses" is 1,839, including 774 neutral, 320 anger, 392 sadness, and 353 happy labeled utterances.

Dialogue Context	
System-1	Hey, why don't you go for a jog? You haven't done much exercise recently. (君、運動不足君だから外でジョギングしようよ。)
User-1	No, I'll be tired. (えー、疲れるからいやだなー。)
Target response	
System-2 (Neutral)	But you will be fat if you have less exercise. (でもね、君、体を動かさないと太っちゃうよ)
Response variations in different emotions	
System-2' (Angry)	Less exercise makes you obese. (でも体を動かさないと太っちゃうでしょ)
System-2' (Sad)	But you will be fatter if you have less exercise... Don't you mind that? (でも…君は体を動かさないともっと太っちゃうよ…それでもいいの？)
System-2' (Happy)	You can solve the problem with your tiredness! (疲れるということは運動不足が解消されるということですね！)

Table 2: An example of collected data. "Dialogue contexts" show the precedent utterances to the target response. "Target response" indicates the target system response to be paraphrased, with its emotion annotation. "Response variations in different emotions" show response variations collected in this work, which have the same meaning as the original "target response" in different emotional expressions. The original corpus was collected in Japanese; thus, the English is a translation.

We used crowd-sourcing to collect response variations in different emotions. Crowdsourcing is a widely used approach for collecting paraphrasing expressions in existing works (Burrows et al., 2013) to cover lexical divergence (Xu et al., 2014; Jiang et al., 2017). In this work, we focus on collecting emotional variations of system utterances. We showed the dialogue context, target response, and a new emotion label, and requested the crowd-sourcing participants to paraphrase the target response with the given new emotion label. In the example in Table 2, the dialogue context and target response are "system-1", "user-1", and "system-2", and the target emotion label is one of the emotions except "neutral": "angry", "sad", and "happy". During the annotation, we showed the participants a figure of Russell ’ s circumplex model (Figure 1) and the following instructions.

1. The response is appropriate to the given dialogue context.
2. The response is expressive to show the given emotion label.
3. The purpose of the system is to persuade the user.

[1] It is possible to express such silence with voice acting; however, it will cause other problems such as turn-taking with human users.

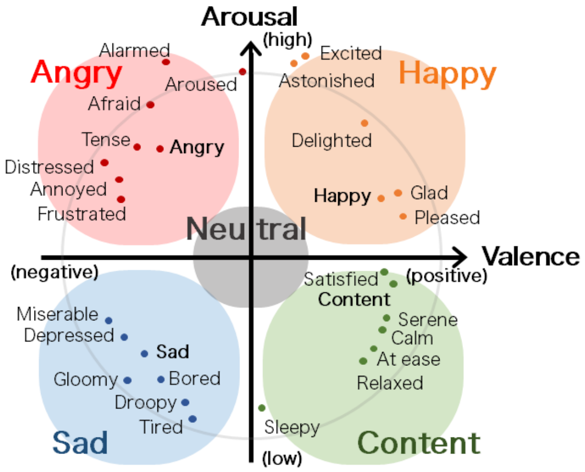

Figure 1: Russell's circumplex model.

Finally, we collected 5,517 variations to 1,839 target responses. The resultant corpus consists of 7,356 response variations to 1,839 dialogue contexts; each cotext has four responses based on different emotion labels.

3. Emotional Speech Recording

It is difficult to express and transfer one's emotion correctly using only a text. Thus, we collected emotional speeches corresponding to the given emotion label of the target system response to improve the expressiveness of the emotion by the system. We recorded speeches from a student of a voice actor school, who is training to be a professional voice actor. It is challenging to record the whole of the collected system's responses (7,356 responses); thus, we ranked each dialogue context. We used BERT to convert a dialogue context to a vector in a latent space and used K-means clustering to select points in the latent space. We selected a variety of dialogue contexts for building a robust dialogue system.

3.1. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT (Devlin et al., 2019) is a language representation model, which is trained from large-scale text corpus with its transformer architecture (Vaswani et al., 2017). The pre-trained model of BERT is trained to predict masked next or previous sentence in language modeling task for making representation of sentences. It is reported that the model well represents a sentence meaning and outperformed existing word or sentence representation methods in several language understanding benchmark tasks. We used a pre-trained model that is trained from Japanese texts on social network services (Sakaki et al., 2019). We input a sentence in our data to the pre-trained model and converted the sentence to a fixed length vector, 768 dimensions.

3.2. Representative Point Selection

As indicated in Eqn (1) and (2), the dialogue systems select or generate response r^t given a dialogue context q^t. This means that the diversity of pairs contained in the training data or selection pool ($< q_i, r_i >$) decides the system robustness.

K	500	550	600	650
overlapping samples	-	300	750	930
total samples	500	750	930	1075

Table 3: Number of selected samples (total samples) and overlapping samples with previous Ks by increasing the number of K from 500 to 650 at 50 intervals.

In this paper, thus, we converted any user utterances (User-1 in the example of Table 2; u^t) and its antecedent system utterance (System-1; r^{t-1}) to vectors $\mathbf{u}^t$ and $\mathbf{r}^{t-1}$ by using BERT. The concatenated vector $\mathbf{q}^t = \mathbf{u}^t \oplus \mathbf{r}^{t-1}$ is used as a point of the dialogue context in the latent space of BERT, as shown in Figure 2.

We applied K-means clustering (Hartigan and Wong, 1979) to select samples to be used for recording. A sample close to the centroid of each cluster is used as a representative sample. For building a robust system, it is expected to select representative samples in several conditions. If we change number of classes K, some centroids are selected in both class numbers, however, some new samples will be selected as shown in Figure 5. In the example of Figure 5, sample number 2 and 3 are such representative samples for both K=4 and 6, but sample number 1 and 4 are not selected in K=6. We tried K = 500, 550, 600 and 650 to select such representative centroids as shown in Table 3. Finally 1075 samples, to be used for the recording, were selected. We eliminated 5 samples that have no overlapping with other Ks, according to the number of samples in their clusters. As a result, we recorded 4,280 emotional speech utterances produced by the voice actor, which explicitly express annotated emotion labels (1,070 dialogue contexts × 4 emotions $=$ 4,280 sample responses).

3.3. Recording Procedure

We recorded emotional speeches according to their labels, and these speeches were spoken by a voice actor. During the recording, we showed a dialogue context (System-1, User-1), a response variation, and attached an annotated emotion label to the variation. The voice actor says the response variations according to the dialogue context and the emotional state. We showed a picture from Figure 3 that corresponds to the emotion label of the current response variation for indicating the emotional state. We also showed emotional gesture samples implemented in communication robot CommU, as shown in Figure 4, which will be the robot platform of our persuasive dialogue system before the recording. We explained to the actor that the recorded voices would be used as robot voices.

4,280 response variations selected by the representative point selection (Section 3.2.) are used for the recording. The total duration of the recorded speeches for each emotion label is shown in Table 4. As shown in the figure, each emotion label has a different duration. The average duration of the "angry" class was shorter because the emotion can be expressed in short, strong words. On the other hand, the emotion class "sad" and "happy" require a longer du-

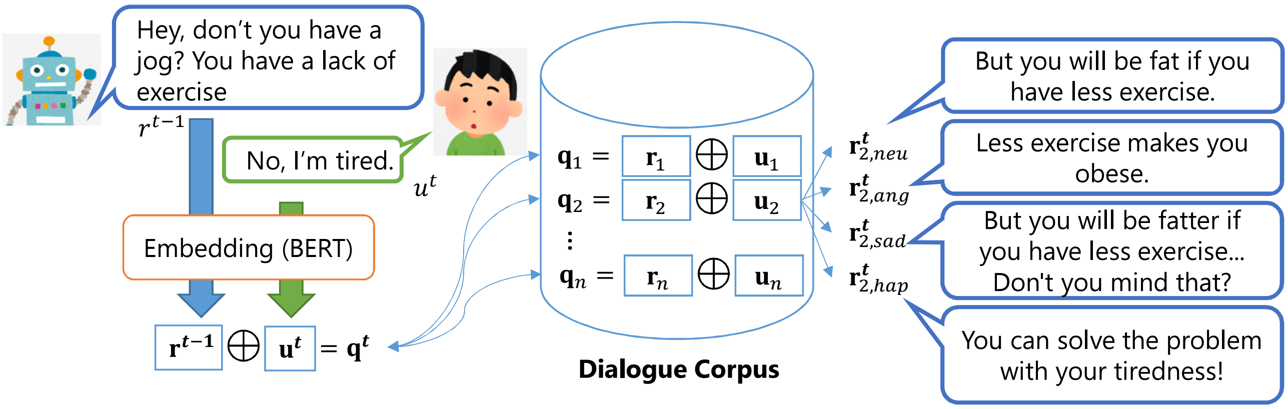
Hey, don't you have a jog? You have a lack of exercise
r^{t-1}
No, I'm tired.
u^t
Embedding (BERT)
$\mathbf{r}^{t-1} \oplus \mathbf{u}^t = \mathbf{q}^t$
$\mathbf{q}_1 = \mathbf{r}_1 \oplus \mathbf{u}_1$
$\mathbf{q}_2 = \mathbf{r}_2 \oplus \mathbf{u}_2$
$\vdots$
$\mathbf{q}_n = \mathbf{r}_n \oplus \mathbf{u}_n$
Dialogue Corpus
$\mathbf{r}^t_{2,neu}$
$\mathbf{r}^t_{2,ang}$
$\mathbf{r}^t_{2,sad}$
$\mathbf{r}^t_{2,hap}$
But you will be fat if you have less exercise.
Less exercise makes you obese.
But you will be fatter if you have less exercise... Don't you mind that?
You can solve the problem with your tiredness!

Figure 2: Persuasive dialogue system based on multi-emotional response corpus.

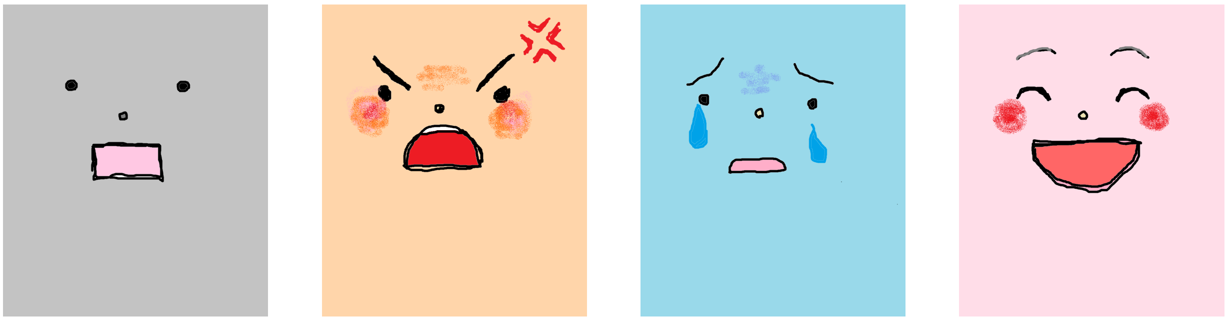

Figure 3: Figures for emotion instruction to the voice actor (neutral, angry, sad and happy from left to right).

Emotion	Length
Neutral	1:04:53.4
Angry	1:03:16.3
Sad	1:19:42.5
Happy	1:09:38.1

Table 4: Recorded speech duration of each emotion class.

ration. We assume that these emotions require some explanations to show the emotion clearly in texts. Before the recording, we had a trial recording of around 100 sentences to stabilize the emotion expression of the actor.

	Text	Speech
Neutral	0.487	0.800
Angry	0.413	0.833
Sad	0.427	0.917
Happy	0.407	0.837
Total	0.433	0.847

Table 5: Results of human evaluations to predict annotated emotion labels.

	Text	Speech
Neutral	0.510	0.840
Angry	0.410	0.900
Sad	0.430	0.950
Happy	0.430	0.850
Total	0.445	0.885

Table 6: Results of human evaluations to predict annotated emotion labels (majority voting).

4. Evaluation on Expressiveness

4.1. Human Subjective Evaluation

We conducted a human subjective evaluation for evaluating the emotional expressiveness of the collected texts and speeches because the expressiveness is important for the persuasive dialogue system as it uses emotional states explicitly. We randomly extracted 100 samples of system response variations from each emotion class as the test-set. We used the test set for human subjective evaluation of the emotional expressiveness of the collected corpus to investigate whether the subjects can predict the original emotional state of the shown sample in text or speech. We assigned three subjects for each sample. The subjects select an emotion class of the shown sample from the “neutral”, “angry”, “sad” and “happy” classes. Table 5 shows ratios that subjects can predict the original labels annotated on test-set samples. Table 6 shows the success ratios as well, but it indicates the majority voting results. In other words, the table shows the ratios of samples for which two or three subjects succeeded in predicting the annotated labels.

These results indicate that speech has better emotion expressiveness than text. 88.5% of the samples are predicted correctly in the majority voting cases. It is clarified that the corpus constructed in this work has higher emotional

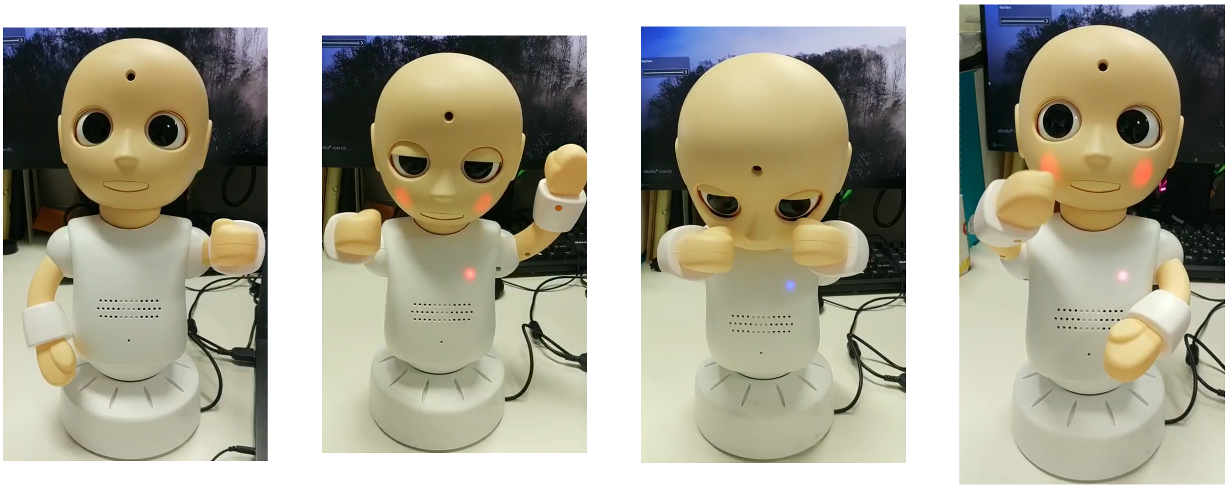

Figure 4: Robot actions (neutral, angry, sad and happy from left to right).

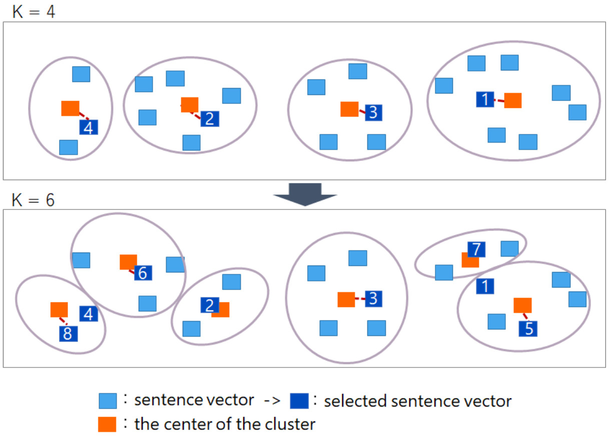
K = 4
4
2
3
1
K = 6
6
4
8
2
3
7
1
5
: sentence vector -> : selected sentence vector
: the center of the cluster

Figure 5: Sample selection based on K-means clustering.

expressiveness in the human subjective evaluation.

4.2. Analysis on Collected Texts

The texts in the collected speech corpora have lower emotion expressiveness; even the speech data has very high expressiveness. We analyzed the text data by classifying it into two categories: “original” (emotional response samples extracted from the original persuasive dialogue corpora) and “collected” (samples newly collected through our data collection based on paraphrasing). Table 7 shows the number of samples for each category in the test set and for each emotion label. We show the detailed results of the human subjective evaluation for the text data in Table 8 and Table 9, as the results for accuracy and majority voting, respectively.

The “neutral” responses in the “original” category have higher emotional expressiveness than the responses in the “collected” category. However, the “collected” category has higher emotional expressiveness than the “original” category for other emotion classes because our data collection is based on a paraphrasing task, which can emphasize the emotional expression. On the other hand, it is indicated that conversion from emotional texts to “neutral” texts is more complicated than conversion from “neutral” to other emotions.

	Original	Collected
Neutral	43	57
Angry	15	85
Sad	19	81
Happy	23	77
Total	100	100

Table 7: Numbers of samples in original and collected categories on each emotion.

	Original	Collected
Neutral	0.550	0.439
Angry	0.378	0.420
Sad	0.281	0.461
Happy	0.348	0.424
Total	0.427	0.436

Table 8: Results of human evaluations to predict annotated emotion labels by giving texts for each categories.

5. Related works

There are many existing works of emotional corpora. Interspeech emotion challenges raised problems to use emotional speeches and proposed general speech features for emotion recognition (Schuller et al., 2009; Schuller et al., 2013). Other types of modalities are also considered in existing works, such as facial expressions of users (Zhang et al., 2016) or combining a variety of modalities (Kaya and Salah, 2016). Most of these works focused on recognition and utilizing user emotions; however, limited numbers of works focused on the system’s emotional expressions. Watanabe et al. (Watanabe et al., 2018) investigated that negative emotion expressions by the android have relations to the user’s decision in the persuasive scenario; however, this work is based on handcrafted scenarios. We proposed to build a dialogue corpus in a persuasive scenario with the system’s emotional expressions in texts

	Original	Collected
Neutral	0.628	0.421
Angry	0.333	0.424
Sad	0.316	0.457
Happy	0.391	0.442
Total	0.470	0.437

Table 9: Results of human evaluations to predict annotated emotion labels by giving texts for each categories. (majority voting).

(Yoshino et al., 2018). This work focuses on emphasizing the system's ability to express emotion by using a speech-based corpus for realizing a practical approach to users in dialogue.

6. Conclusion

In this paper, we explained our collected emotional speech corpus, which is constructed for a persuasive dialogue system with emotional states and expressions. We extended the existing persuasive dialogue corpora with emotional expressions as a multi-emotional response corpus, including recorded emotional speech. We defined the paraphrasing task on crowd-sourcing to extend the text corpus for getting a variety of responses with given emotion labels. Emotional speeches by a voice actor were recorded to improve the expressiveness of the dialogue system. We used BERT and K-means clustering for selecting sub-dialogue samples to be used for recording, covering diverse dialogue contexts. We evaluated the emotional expressiveness of the collected texts and speeches in human subjective evaluation. We showed that the collected emotional speeches have high emotion expressiveness (88.5% in majority voting). In our analysis, it was investigated that the defined emotion paraphrasing task by using crowd-sourcing can collect more expressive response variations for most emotional states apart from "neutral".

As our future work, we need to evaluate the coverage of the collected samples by applying to a persuasive dialogue system. We will build a persuasive dialogue robot with collected emotional speeches for improving the persuasion success rate.

7. References

Adler, R. F., Iacobelli, F., and Gutstein, Y. (2016). Are you convinced? a wizard of oz study to test emotional vs. rational persuasion strategies in dialogues. *Computers in Human Behavior*, 57:75–81.

Burrows, S., Potthast, M., and Stein, B. (2013). Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):1–21.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Fogg, B. J. (1999). Persuasive technologies. *Communications of the ACM*, 42(5):27–29.

Forgas, J. P. (1998). On feeling good and getting your way: Mood effects on negotiator cognition and bargaining strategies. *Journal of personality and social psychology*, 74(3):565.

Georgila, K. (2013). Reinforcement learning of two-issue negotiation dialogue policies. In *Proceedings of the SIGDIAL 2013 Conference*, pages 112–116.

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., and Scherer, S. (2017). Affect-lm: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hiraoka, T., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2016). Learning cooperative persuasive dialogue policies using framing. *Speech Communication*, 84:83–96.

Jiang, Y., Kummerfeld, J. K., and Lasecki, W. (2017). Understanding task design trade-offs in crowdsourced paraphrase collection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 103–109.

Kaya, H. and Salah, A. A. (2016). Combining modality-specific extreme learning machines for emotion recognition in the wild. *Journal on Multimodal User Interfaces*, 10(2):139–149.

Keltner, D. and Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition & Emotion*, 13(5):505–521.

Mazzotta, I., de Rosis, F., and Carofiglio, V. (2007). Portia: A user-adapted persuasion system in the healthy-eating domain. *IEEE Intelligent systems*, 22(6):42–51.

Morris, M. W. and Keltner, D. (2000). How emotions work: The social functions of emotional expression in negotiations. *Research in organizational behavior*, 22:1–50.

Russell, J. A. (1978). Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152.

Sakaki, T., Mizuki, S., and Gunji, N. (2019). BERT pre-trained model trained on large-scale Japanese social media corpus.

Schuller, B., Steidl, S., and Batliner, A. (2009). The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association.*

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the In-*

ternational Speech Communication Association, Lyon, France.

Sinaceur, M. and Tiedens, L. Z. (2006). Get mad and get more than even: When and why anger expression is effective in negotiations. *Journal of Experimental Social Psychology*, 42(3):314–322.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., and Yu, Z. (2019). Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *ACL*.

Watanabe, M., Ogawa, K., and Ishiguro, H. (2018). At the department store can androids be a social entity in the real world? In *Geminoid Studies*, pages 423–427. Springer.

Xu, W., Ritter, A., Callison-Burch, C., Dolan, W. B., and Ji, Y. (2014). Extracting lexically divergent paraphrases from twitter. *Transactions of the Association for Computational Linguistics*, 2:435–448.

Yoshino, K., Ishikawa, Y., Mizukami, M., Suzuki, Y., Sakti, S., and Nakamura, S. (2018). Dialogue scenario collection of persuasive dialogue with emotional expressions via crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zhang, Z., Girard, J. M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., et al. (2016). Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446.

Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.