

Dialogue Scenario Collection of Persuasive Dialogue with Emotional Expressions via Crowdsourcing

Koichiro Yoshino^{1,2}, Yoko Ishikawa¹, Masahiro Mizukami¹,
Yu Suzuki¹, Sakti Sakriani¹, and Satoshi Nakamura¹

¹Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5, Takayama-cho, Ikoma, 6300192, Japan

²PRESTO, Japan Science and Technology Agency, 4-1-8 Hon-machi, Kawaguchi, Saitama 332-0012, Japan
{koichiro, ishikawa.yoko.io5, masahiro-mi, ssakti, ysuzuki, s-nakamura}@is.naist.jp

Abstract

Existing dialogue data collection methods such as the Wizard of Oz method (WoZ) or real dialogue recording are costly, and they prevent launching a new dialogue system. In this study, we requested crowd workers in crowdsourcing to create dialogue scenarios according to the instruction of the situation for persuasive dialogue systems that use emotional expressions. We collected 200 dialogues in 5 scenarios for a total of 1,000 via crowdsourcing. We also annotated emotional states and users' acceptance for system persuasion by using crowdsourcing. We constructed a persuasive dialogue system with the collected data and evaluated the system by interacting with crowd works. From the experiment, it was investigated that the collected labels have sufficient agreement even if we did not impose any training of annotation to workers.

Keywords: Dialogue corpus, emotion labels, persuasive dialogue, crowdsourcing

1. Introduction

Expressing one's emotion or feeling with language helps us to understand each other in human-human communication. We cannot observe the internal states of other people such as our emotions or feelings directly, and we indirectly estimate the internal states from observable states such as utterances and the behaviors of others. Explicitly expressing internal states helps to build a good relationship between humans by reducing misunderstandings and unrequited feelings (Ekman, 1993). For example, sharing the same emotional state helps us to develop close relationships with others.

Expressing emotional states is very effective to process the persuasion or negotiation (Keltner and Haidt, 1999; Morris and Keltner, 2000). Persuaders who express positive emotions increase the ratio of success at persuasion more than persuaders who do not express their emotions because expressing positive emotions gives a cooperative impression to a partner (Carnevale and Isen, 1986; Forgas, 1998). On the other hand, expressing negative emotions such as "anger" may wrest a concession from a partner even if the proposal from the persuader is not attractive to the partner, especially when the partner does not have any other options to choose (Sinaceur and Tiedens, 2006). Using emotional expressions is one of the most effective ways to change the belief or behaviors of a partner to increase the success rate of persuasion or negotiation.

Captology (Computers as persuasive technologies) (Fogg, 1997) is a research area to work on systems that affect the beliefs or behaviors of users. A persuasive dialogue system is known as a part of this research area, and some dialogue systems have been developed that can change the actions or behaviors of users with persuasion. Some studies of persuasive dialogue systems investigated efficient dialogue strategy on dialogue management to persuade users. Hiraoka et al., (2016) introduced actions of framing and logical explanations of advantages and disadvantages of products for

persuading the users to purchase the product.

The major problem of implementing a dialogue system is data, in any domains or tasks of systems, because most methods of dialogue modeling are based on statistical methods that require large-scale data-sets. However, collecting new dialogue data in accordance with a defined new task is costly. Some approaches enable easy data collection in a new domain by utilizing Web data (Banchs and Li, 2012) or by extracting dialogue parts from chat-like conversations (Nio et al., 2014). However, collecting large-scale dialogue data that contain emotional expressions are still difficult. Emotional expressions tend to be observed in communications between people who have close relationships. However, it is hard to record dialogues in such closed situations. It is also difficult to extract such conversations from Web because expressing emotions in public space is somewhat suppressed.

Crowdsourcing has attracted attention as an efficient way for collecting or expanding dialogue data (Yu et al., 2016). We collected dialogue data which contains emotional expressions by requesting crowd workers to create complete dialogue scenarios. This approach makes it possible to create a dialogue scenario even if it is hard to record the dialogue data in the hypothetical situation of the scenario. One large concern in this approach is whether we can collect realistic dialogue data that can be used for the training of dialogue systems. We built and evaluated a simple persuasive dialogue system based on an example-based approach to investigate that the collected data is usable for the training of the dialogue system. The resultant model worked well even if the dialogue scenario collection of one dialogue was conducted by a single annotator. In this work, we used crowdsourcing not only for dialogue scenario collection but also for label annotations. We also found that the labels had sufficient quality for them to be used for system construction, even if they were annotated by crowd workers.

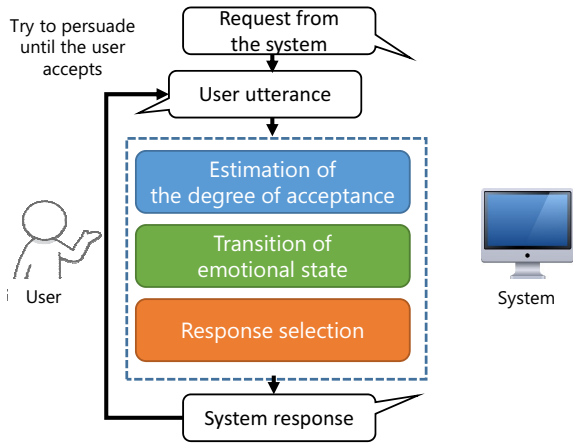


Figure 1: The overview of the proposed persuasive dialogue system with emotional expressions.

2. Persuasive Dialogue Scenario

This section describes the scenario of persuasive dialogue systems that we assumed in our data collection. The flow of persuasion from the system is shown in **Figure 1**. A persuasive dialogue system talks with users about their living habits and tries to change these habits. The dialogue starts with a request from the system to the users, and the system tries to continue with persuasion until the users accept the request. The emotional states and the degrees of the users' acceptance are estimated by the system. The dialogue ends if the system recognizes the acceptance of the users or if a pre-defined number of turns pass (=failure). In this scenario, the dialogue data of the persuasive dialogue labeled with emotional states and the degree of the users' acceptance are required. The collection procedure of these data, the dialogue scenarios, annotations of emotion, and annotations of the degree of users' acceptance are described in the following sections.

3. Emotion Labeled Corpus Construction Through Crowdsourcing

A labeled corpus of persuasive dialogue is required to construct the persuasive dialogue system. We collected an emotionally persuasive corpus through crowdsourcing (Howe, 2006). In the data collection, first, we requested that crowd workers write dialogue scenarios of persuasive dialogue by using emotional expressions. Then, labels of the users' acceptance and emotion were annotated by other crowd workers.

3.1. Scenarios Collection

To collect the scenarios, we requested that crowd workers write dialogue scenarios of persuasive dialogue in accordance with the following instructions.

- The dialogue starts with a system suggestion for changing daily activities
- The system tries to persuade a user with some emotional expressions
- The dialogue ends with the user accepting the suggestion

- The dialogue consists of more than 20 utterances

The specific suggestions to the crowd workers are as follows.

Instructions given to crowd workers

John is living with a robot who guides his daily life to improve his living habits. The robot encourages John to change his habit if he finds a bad habit. For example, when the robot find that John does not get enough exercise, he presses John to go jogging. However, John does not listen his advice even if the robot describes the reason which John should follow his advice. Thus, the robot decided to persuade John by using emotional expressions including anger, sadness, and happiness.

Create a dialogue example of the persuasion between John and the robot. The dialogue starts with a system suggestion for getting exercise. The robot tries to persuade John by using various emotional expressions, and finally, John accepts the offer of the robot. The robot must use one or more emotional expressions from emotion categories of happiness, sadness, and anger. The dialogue consists of more than 20 utterances.

We prepared five daily-life guidance scenarios: "Clean the room (cleaning)," "Don't leave a dish unfinished (lunch)," "Sleep early (sleep)," "Stop playing the game (game)" and "Get some exercise (exercise)."

Table 1: Numbers of scenarios and utterances

Scenarios	Dialogues	$\#q_i$	$\#r_i$
Cleaning	200	2282	2292
Lunch	200	2173	2185
Sleep	200	2147	2180
Game	200	2175	2155
Exercise	200	2203	2216

Table 1 shows the number of collected dialogues and utterances for each scenario. A total of 1000 dialogues (200 dialogues for each scenario) were collected with crowdsourcing. One crowd worker generated 5 dialogues at most (1 for each scenario). We annotated labels for every dialogue to use as the training data of the dialogue system.

3.2. Emotional States

We used five emotional states: "neutral," "happy," "contented," "angry," and "sad," which are defined in Russell's Circumplex Model (Russell, 1978). Two dimensions were used: valence and arousal, in Russell's model, and the emotional state was decided using degrees of these two dimensions. States were labeled as "Happy," "Angry," "Sad" and "Contented" from the first quadrant to the fourth quadrant. Around the origin was annotated with "Neutral."

3.3. Degree of Users' Acceptance

Knowing the degree of users' acceptance, that is, whether or not the users will accept the system's suggestions, is necessary to construct a persuasive dialogue system because

the degree is used for the detection of the end of the dialogue with the acceptance of the users. We defined the users' acceptance in 5 degrees (1: Refused, 3: Not sure, 5: Accepted).

3.4. Label Annotation

We made annotations of emotion labels and users' acceptance rates for persuasion with 5 degrees to construct a persuasive dialogue system with emotional expressions. We requested that crowd workers annotate both the degree of users' acceptance for the user utterances and the emotional labels for each utterance. Three annotators were assigned for one label, and we utilized labels that had 2 or 3 agreements. We removed examples that had 3 different labels by 3 annotators, because it probably be caused by the difficulty of the annotation or qualities of crowd workers assigned for the example. Users' acceptances in 5 degrees (5: accept, 4: possibly accept, 3: cannot say, 2: possibly reject and 1: reject) were annotated on user queries in dialogues, and emotional states with 5 kinds of labels were annotated on every user query and system response. The crowd workers read every utterance of the dialogue scenario before the annotation and annotate the acceptance and the emotion label for each utterance. The figure of Russell's Circumplex Model (Russell, 1978) is presented crowd workers before every annotation session. The figure includes some example expressions of each emotion classes (e.g. glad or enjoyable belong to the "happy" class). Table 2 Table 3 and Table 4 show the percentage of annotated labels for each kind of annotation, and "None" means no label was assigned for the example because the annotation for the example was divided.

In this corpus, 30% of the utterances were annotated as "Neutral," and more than 20% utterances were annotated as "Angry" or "Sad" for each label. In details, the proportions of negative emotions ("Angry" and "Sad") in system utterances were smaller than the proportions of negative emotions in user utterances. It indicates that created scenarios contain some examples to elicit positive emotion for negative user utterances by using positive emotions. The proportions of negative emotions of users are significant because we instructed scenario writers to write more extended scenarios (more than 20 utterances) by the final agreement of the system and the user. The proportion of "None" of system utterances was little small because scenario writers are conscious of using emotional expressions on system utterances. Table 4 shows an example of an annotated corpus that we collected.

3.5. Annotation Agreement

As results, we collected 1,000 persuasive dialogue corpus annotated with emotional states and users' acceptance rate for the persuasion. To examine the quality of the collected corpus via crowdsourcing, we calculated Fleiss' Kappa value (Fleiss et al., 2013), the rate of concordance of each annotator, for both the degree of users' acceptance and the emotional label. Fleiss' Kappa of the emotion labels without removing disagreed examples was 0.345. After the removing of disagreed examples as described in Section 3.4., the Fleiss' Kappa was 0.411, which means moderate agree-

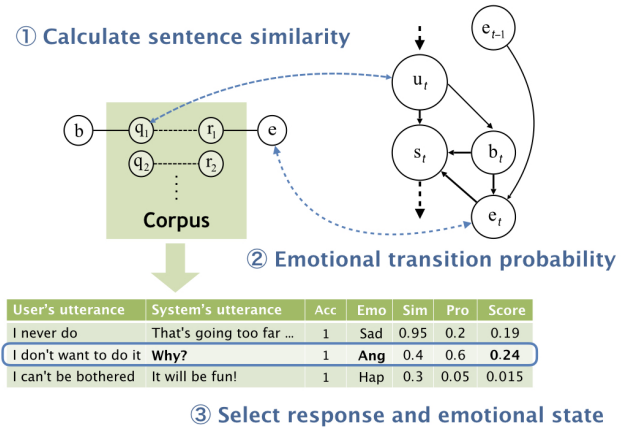


Figure 2: Response selection of dialogue system with emotional states

ment. Fleiss' Kappa of the degree of users' acceptance was 0.370, which means fair agreement. We also calculated the mean squared error value of the degree of users' acceptance with the following equation.

$$\frac{|x_A - x_B|^2 + |x_B - x_C|^2 + |x_C - x_A|^2}{3}. \quad (1)$$

Here, $A-C$ are IDs of annotators. The mean squared error value was 0.850 (lower is better), which is low enough for 5-degree annotations.

The Kappa value of emotion labels was over 0.4, and the Kappa value of the users' acceptance was lower than 0.4. However, the mean square error of the user's acceptance was less than 1.0, and it indicated the collected data has fair agreement to be used as the training data of the system. We removed examples that had three different labels by three annotators from the training data of the system; thus, the quality of the data used for the system training is much better than this score. We have not compared the annotation results by crowd workers with any annotation results of well-trained annotators. An existing work reported 0.52 Kappa value for valence-arousal emotion labeling by using audio and video data (Konar and Chakraborty, 2015). Our work only uses text data to decide emotion labels; thus, our Kappa value is not particularly low even if we used crowd workers for annotations. However, comparing results by well-trained annotator and crowd workers in the same setting is still a remaining future work.

Collected data included many utterances that had negative emotion labels such as "Angry" and "Sad," because the instruction to crowd workers focused on the process of persuasion. Positive emotions only happened at the last part of the dialogue; thus, the number is smaller than the number of negative emotions.

4. Persuasive Dialogue System Trained from The Collected Data

We built a persuasive dialogue system based on the example-based architecture according to the belief-desire theory of emotion (Reisenzein, 2009). In this theory, the combination of a desire and a belief evokes emotions. The desire means the goal to be achieved, and the belief implies

Table 2: Instructions for users' acceptance

Label	Label meaning	Points to be checked
1	John rejects the suggestion of the robot	Does the sentence contains expressions of strong rejection?
2	John possibly reject the suggestion of the robot	Does the sentence contains expressions of hesitate?
3	It cannot be said John accepts or rejects the suggestion of the robot	Does the sentence contains ambiguous expression? Is it difficult to find any rejecting or accepting expressions?
4	John possibly accept the suggestion of the robot	Does the sentence contains expressions of agreement?
5	John accepts the suggestion of the robot	Does the sentence contains expressions of clear acceptance?

Table 3: Percentage of acceptance labels in corpus

label	percentage
1	23.36%
2	26.39%
3	17.211%
4	10.27%
5	9.78%
NONE	9.99%

Table 4: Percentage of emotion labels in corpus

Label	Total	User	System
Neutral	28.73%	22.93%	33.54%
Happy	8.74%	6.91%	10.57%
Contentment	4.18%	5.21%	3.15%
Angry	23.04%	23.36%	22.72%
Sad	24.90%	28.89%	20.91%
None	10.41%	11.46%	9.36%

the belief of environment observation including achievement of the desired goal. The belief-desire model evokes positive emotions if the belief is approaching the desired goal. However, the model evokes negative emotions if the belief is leaving the desired goal. For example, if someone wants to go to picnic (=desire) but it is difficult to go to a picnic because the weather condition is bad (=belief), negative emotions are evoked because the desire may not be achieved. On the other hand, if there are small mismatches between the desire and the belief, the model evokes a positive emotion such as "happy" or "contentment". For example, if someone wants to go to picnic (=desire) and it is possible because the weather condition is good (=belief), a positive emotion is evoked.

In the persuasive scenario, we can use the users' acceptance and success of persuasion as the belief and desire of the system. The system transits its own emotional state: if the degree of user's acceptance is low, the system transits the emotional state to "anger" or "sad", and if the degree of user's acceptance is high, the system transits the emotional state to "happy", or "contentment". In the proposed architecture, the system selects a response \hat{s}_t and its emotional state \hat{e}_t by using the given user utterance u_t and estimated belief (user's acceptance) b_t (Figure 2).

The belief estimation (estimation of users' acceptance) is modeled by Support Vector Regression (SVR). SVR is an expansion of Support Vector Machine (SVM) for the re-

Table 5: Example of annotated corpus

Speaker	Utterance	Accept	Emotion
Sys	Let's clean the room	—	Neutral
User	No, it's a bother.	1	Sad
Sys	I see it. There is some trash on the floor	—	Neutral
User	I'll do it later. Not now.	NONE	Sad
Sys	You always say you'll do it but end up doing nothing.	—	Angry
User	Sure.	3	Sad
Sys	You don't have to say you'll do it. I'll support you regardless.	—	Happy
User	OK, I do.	5	Sad
Sys	Great, let's give it all you've got!	—	Happy

Table 6: Agreements of annotators

	emotion	acceptance
#annotations	22,008	10,980
Fleiss' Kappa	0.411	0.370
Mean square error	—	0.850

gression problem. SVR has high generalizing capability because the learning of SVR minimizes the upper bound of generalization error. We used the corpus that is annotated with the degree of user's acceptance as described in Section 3.4.. To make feature vectors from the user's utterances for the regression, we extracted words as linguistic features from the user utterance by using morphological analyzer Mecab (Kudo et al., 2004). Synonyms of words in the user utterance are extracted by using WordNet (Bond et al., 2012) to extend the word feature vector. We also used positive/negative score by using pre-defined dictionary of positive/negative words (Takamura et al., 2005). Each extracted vector is concatenated as a single vector to be used as the input of SVR. The regression learns the annotated degree of user's acceptance for each utterance.

The system calculates the cosine similarity $\cos(u_t, q_j)$ for each pair of the user utterance u_t and a user-query in the example database q_j (Figure 2-①). The example database consists of pairs of a user query q_j and its response r_j with annotations of the user's acceptance of the query b_j and the emotional state of the response e_j . The database consist of query-response pairs that are extracted from the collected corpora.

In the step of Figure 2-②, the system calculates the posterior probability of emotion e_j with $P[e_j]$ that is determined by the transition probabilities $P(e_t|e_{t-1}, b_t)$ estimated with maximum likelihood estimation on the training dialogue corpus.

At the last step of Figure 2-③, the system gives a score for each query-response pair $\langle q_i, r_i \rangle$ as,

$$\text{score}(\langle q_j, r_j \rangle) = \cos(u_t, q_j) \times P[e_j]. \quad (2)$$

Finally the system responds with the \hat{r}_t that has the highest score. Each response r_j has the annotation of the emotion label e_j ; thus, the emotion of the system \hat{e}_t is also decided by the score.

We evaluated the dialogue system through real dialogue with crowd workers on crowdsourcing. Through crowdsourcing, 92 users including 57 females and 35 males participated in the evaluation. Crowd workers actually talked with dialogue systems in different five scenarios we defined (cleaning, lunch, sleep, game and exercise). Each dialogue system tried to persuade the worker until the worker accepted the suggestion of the system. The order of the system was randomly selected. The dialogue was ended after the worker accepted the request or after 20 dialogue turns passed (=failure). After the dialogue, workers answered four questions for each dialogue system with a 5-level subjective score (1: Disagree, 3: Not sure, 5: Agree). The four questions were defined as:

- **NATURALNESS:** Did you feel that the system’s response was natural?
- **PERSUASIVENESS:** Did you feel that the system’s suggestion was persuasive?
- **KINDNESS:** Did you feel that the system kindly talked with you?
- **HUMANLIKENESS:** Did you feel that the system was humanlike?

As the baseline system, we also constructed a system that does not have emotion transition architecture. The baseline system select a response in example-based dialogue manner from response candidates that is annotated with “neutral” emotion label.

Table 7: Average scores of subjective evaluation.

Question	w.o. emotion	w. emotion
NATURALNESS	3.402	3.293
PERSUASIVENESS	3.598	3.522
KINDNESS	3.511	3.576
HUMANLIKENESS	3.511	3.522

4.1. Evaluation Results

Table 7 shows the results of the subjective evaluation. The evaluation results indicate that the collected corpus had sufficient quality for it to be used for the training data of dialogue systems. The scores for the proposed model with

Table 8: A dialogue example with positive emotional expressions, which is scored high

Turn	Utterance	Accept.	Emo.
Sys 1	Let’s have a jogging tomorrow to solve the lack of fitness.	—	Neutral
Usr 1	I’m not a morning person.	0.80	—
Sys 2	I’ll be happy to attend you.	—	Happy
Usr 2	Really? I will consider to go.	2.82	—

Table 9: A dialogue example with positive emotional expressions, which is scored low

Turn	Utterance	Accept.	Emo.
Sys 1	You should clean-up your room.	—	Neutral
Usr 1	I’ll do that later	0.78	—
Sys 2	You can invite your friends if you clean up the room	—	Happy
Usr 2	I don’t have any friends...	0.84	—

emotional states was not significantly higher than the baseline model without emotional states. However, questionnaires for evaluation participants indicated that some users are positive for the system that has emotional states; thus, we have case studies in the next session.

4.2. Case Studies

Table 8 and 9 show example dialogues of using “Happy” emotions by the system. The first example is scored high, however, the second example is scored low, even if the system uses the same emotion in these examples. Two hypotheses cause the difference: user preference for the system emotion and dialogue context. If the user does not like to talk with emotional people, the dialogue evaluation will be negative even though the system uses appropriate emotional responses. The other problem is caused by the example-based dialogue response selection. The example-based response selection only can consider the contents of the previous user utterance. However, the system still requires additional information to select the appropriate response (e.g., the user does not have any friends).

We show a distribution of human evaluation scores of each metrics to look at the evaluator dependent scores in Figure 3–7. Differences in scores for the system with emotional state and the system without the emotional state (w.emotion-w.o.emotion) are calculated for each evaluator. These results indicate that some users prefer to talk with the system that has an emotional state, in contrast to users who do not like communicating with the system with emotional expressions. Figure 7 indicates that some evaluators who give low scores for some metrics also give low scores for other metrics.

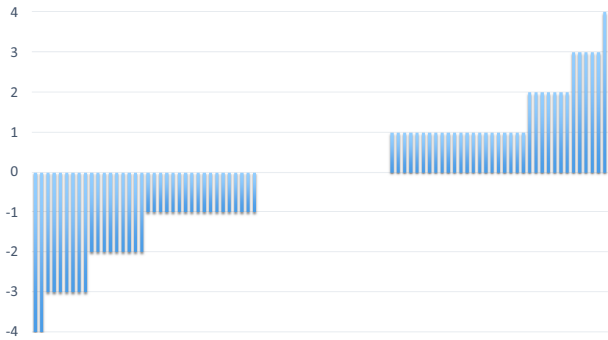


Figure 3: Evaluations of naturalness for each example (w.emotion-w.o.emotion)

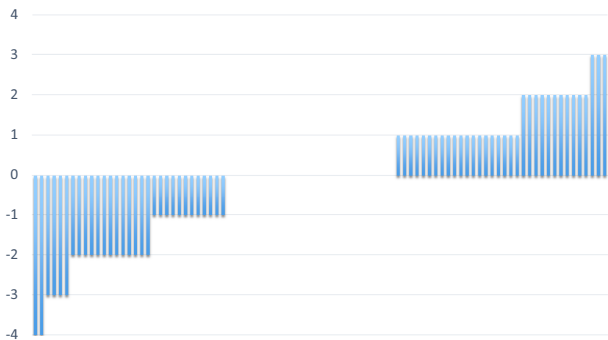


Figure 4: Evaluations of persuasiveness for each example (w.emotion-w.o.emotion)

5. Conclusion

In this study, we constructed a dialogue corpus for persuasive dialogue systems via crowdsourcing, including the labeling of emotional states and the acceptance of users' utterances. Labels of emotions and users' acceptances were also annotated by crowd workers. The labeling results had moderate variance; however, using several annotators contributed to increasing the number of usable labeled utterances for training. We also evaluated the dialogue system trained with the collected dialogue data via crowdsourcing. The results indicated that the corpus has sufficient quality for it to be used as a training set of the dialogue system, even if one crowd worker created a scenario of a conversation for entire dialogue.

6. Acknowledgement

This work is supported by JST PRESTO (JPMJPR165B).

7. References

Banchs, R. E. and Li, H. (2012). Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Bond, F., Baldwin, T., Fothergill, R., and Uchimoto, K. (2012). Japanese semcor: A sense-tagged corpus of japanese. In *Proceedings of the 6th global wordnet conference*, pages 56–63. Global Wordnet Association.

Carnevale, P. J. and Isen, A. M. (1986). The influence of positive affect and visual access on the discovery of integrative solutions in bilateral negotiation. *Organizational behavior and human decision Processes*, 37(1):1–13.

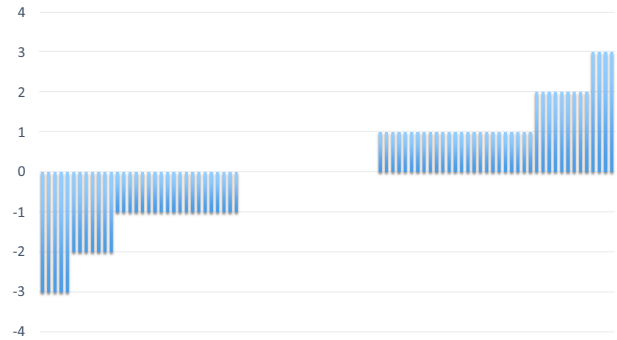


Figure 5: Evaluation of kindness for each example (w.emotion-w.o.emotion)

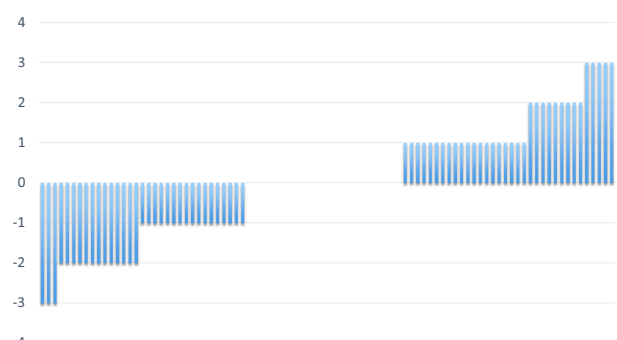


Figure 6: Evaluation of humanlikeness for each example (w.emotion-w.o.emotion)

Ekman, P. (1993). Facial expression and emotion. *American psychologist*, 48(4):384.

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.

Fogg, B. (1997). Captology: the study of computers as persuasive technologies. In *Extended Abstracts on Human Factors in Computing Systems*, pages 129–129. Association for Computing Machinery.

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.

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

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

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

Konar, A. and Chakraborty, A. (2015). *Emotion Recognition: A Pattern Analysis Approach*. Wiley.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to japanese morphological analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 230–237.

Morris, M. W. and Keltner, D. (2000). How emotions work: The social functions of emotional expression

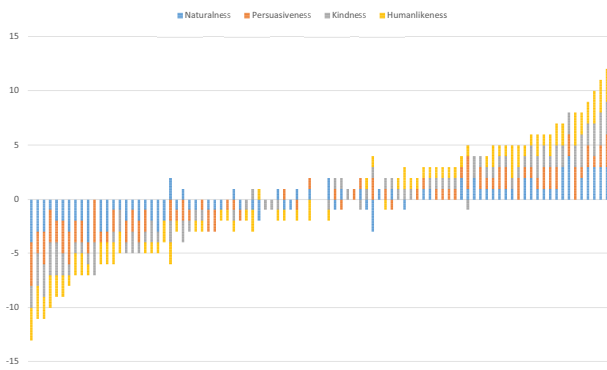


Figure 7: Overall evaluation of each example (w.emotion-w.o.emotion)

- in negotiations. *Research in organizational behavior*, 22:1–50.
- Nio, L., Sakti, S., Neubig, G., Toda, T., and Nakamura, S. (2014). Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system. *IEICE TRANSACTIONS on Information and Systems*, 97(6):1497–1505.
- Reisenzein, R. (2009). Emotions as metarepresentational states of mind: Naturalizing the belief–desire theory of emotion. *Cognitive Systems Research*, 10(1):6–20.
- Russell, J. A. (1978). Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152.
- 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.
- Takamura, H., Inui, T., and Okumura, M. (2005). Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140. Association for Computational Linguistics.
- Yu, Z., Xu, Z., Black, A. W., and Rudnicky, A. (2016). Chatbot evaluation and database expansion via crowdsourcing. In *Proceedings of the RE-WOCHAT workshop of LREC, Portoroz, Slovenia*.