Building a Dialogue Corpus Annotated with Expressed and Experienced Emotions

Tatsuya Ide and **Daisuke Kawahara**
Department of Computer Science and Communications Engineering, Waseda University
{t-ide@toki., dkw@}waseda.jp

Abstract

In communication, a human would recognize the emotion of an interlocutor and respond with an appropriate emotion, such as empathy and comfort. Toward developing a dialogue system with such a human-like ability, we propose a method to build a dialogue corpus annotated with two kinds of emotions. We collect dialogues from Twitter and annotate each utterance with the emotion that a speaker put into the utterance (expressed emotion) and the emotion that a listener felt after listening to the utterance (experienced emotion). We built a dialogue corpus in Japanese using this method, and its statistical analysis revealed the differences between expressed and experienced emotions. We conducted experiments on recognition of the two kinds of emotions. The experimental results indicated the difficulty in recognizing experienced emotions and the effectiveness of multi-task learning of the two kinds of emotions. We hope that the constructed corpus will facilitate the study on emotion recognition in a dialogue and emotion-aware dialogue response generation.

1 Introduction

Text-based communication has become indispensable as society accelerates online. In natural language processing, communication between humans and machines has attracted attention, and the development of dialogue systems has been a hot topic. Through the invention of Transformer (Vaswani et al., 2017) and the success of transfer learning (e.g., Radford et al. (2018); Devlin et al. (2019)), the performance of natural language understanding models and dialogue systems continues to improve. In recent years, there have been studies toward building open-domain neural chatbots that can generate a human-like response (Zhou et al., 2020; Adiwardana et al., 2020; Roller et al., 2021).

One of the keys to building more human-like chatbots is to generate a response that takes into account the emotion of the interlocutor. A human would recognize the emotion of the interlocutor and respond with an appropriate emotion, such as empathy and comfort, or give a response that promotes positive emotion of the interlocutor. Accordingly, developing a chatbot with such a human-like ability (Rashkin et al., 2019; Lubis et al., 2018, 2019) is essential. Although several dialogue corpora with emotion annotation have been proposed, an utterance is annotated only with a speaker's emotion (Li et al., 2017; Hsu et al., 2018) or a dialogue as a whole is annotated (Rashkin et al., 2019), all of which are not appropriate for enabling the above ability.

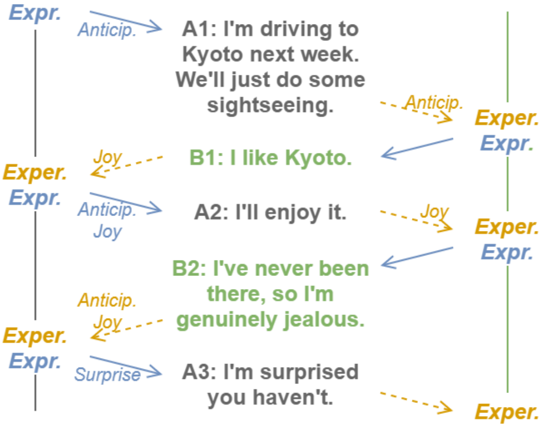

Expr.
Anticip.
A1: I'm driving to Kyoto next week. We'll just do some sightseeing.
Anticip.
Exper.
Expr.
Joy
B1: I like Kyoto.
Exper.
Expr.
Anticip.
Joy
A2: I'll enjoy it.
Joy
Exper.
Expr.
B2: I've never been there, so I'm genuinely jealous.
Anticip.
Joy
Exper.
Expr.
Surprise
A3: I'm surprised you haven't.
Exper.

Figure 1: An example dialogue with expressed and experienced emotions.

In this paper, we propose a method to build an emotion-annotated multi-turn dialogue corpus, which is necessary for developing a dialogue system that can recognize the emotion of an interlocutor and generate a response with an appropriate emotion. We annotate each utterance in a dialogue with an **expressed emotion**, which a speaker put into the utterance, and an **experienced emotion**, which a listener felt when listening to the utterance.

To construct a multi-turn dialogue corpus annotated with these emotions, we collect dialogues

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Student Research Workshop, pages 21 - 30
May 22-27, 2022 ©2022 Association for Computational Linguistics

from Twitter and crowdsource their emotion annotation. As a dialogue corpus, we extract tweet sequences where two people speak alternately. For the emotion annotation, we adopt Plutchik's wheel of emotions (Plutchik, 1980) as emotion labels and ask crowdworkers whether an utterance indicates each emotion label for expressed and experienced emotion categories. Each utterance is allowed to have multiple emotion labels and has an intensity, strong and weak, according to the number of crowdworkers' votes. We build a Japanese dialogue corpus as a testbed in this paper, but our proposed method can be applied to any language.

Using the above method, we constructed a Japanese emotion-tagged dialogue corpus consisting of 3,828 dialogues and 13,806 utterances.[1] Statistical analysis of the constructed corpus revealed the characteristics of words for each emotion and the relationship between expressed and experienced emotions. We further conducted experiments to recognize expressed and experienced emotions using BERT (Devlin et al., 2019). We defined the task of emotion recognition as regression and evaluated BERT's performance using correlation coefficients. The experimental results showed that it was more difficult to infer experienced emotions than expressed emotions, and that multi-task learning of both emotion categories improved the overall performance of emotion recognition. From these results, we can see that expressed and experienced emotions are different, and that it is meaningful to annotate both. We expect that the constructed corpus will facilitate the study on emotion recognition in dialogue and emotion-aware response generation.

2 Related Work

2.1 Emotion-Tagged Corpora

Many non-dialogue corpora annotated with emotions have been constructed. EmoBank (Buechel and Hahn, 2017) is a corpus of social media or reviews with emotion annotation. They annotate sentences with the emotions of a person who read them and a person who wrote them. WRIME (Kajiwara et al., 2021) is an emotion-annotated corpus in Japanese, where SNS posts are tagged with both *subjective* and *objective* emotions. The concept of this corpus is similar to EmoBank. However, they emphasize the subjectivity of annotation and ask writers to annotate their own sentences with emotions. Furthermore, EmoInt (Mohammad and Bravo-Marquez, 2017) aims at the task of detecting emotion intensity. They annotate Twitter posts with anger, fear, joy, and sadness and give each emotion a real value between 0 and 1 as the intensity level.

Some corpora are tagged with non-emotional factors, along with emotions. EmotionStimulus (Ghazi et al., 2015) and GroundedEmotions (Liu et al., 2017) are corpora that focus on the reason for an expressed emotion. The former uses FrameNet to detect a cause, while the latter treats weather and news as external emotion factors. In terms of emotion labels, the two corpora adopt seven emotions (Ekman's six emotions (Ekman, 1992) and shame) and two emotions (only happiness and sadness), respectively. In StoryCommonsense (Rashkin et al., 2018), a series of sentences comprising of a short story is tagged with *motivation* and *emotional reaction* for each character. For emotion labels, they use some theories of psychology, including Plutchik's wheel of emotions (Plutchik, 1980).

None of the above corpora, however, are relevant to dialogue. StoryCommonsense is similar to ours but differs in that characters in a story are annotated instead of speakers' utterances.

2.2 Dialogue Corpora

Several dialogue corpora annotated with emotions are available. DailyDialog (Li et al., 2017) is one collected from educational websites and tagged with emotions and intentions. EmotionLines (Hsu et al., 2018) is a multi-turn dialogue corpus with annotation of emotions. Both of them use seven labels for tagging: Ekman's six emotions (Ekman, 1992) and an other/neutral emotion. MELD (Poria et al., 2019) is an extension of EmotionLines, tagged with not only emotions but also visual and audio modalities. EmpatheticDialogues (Rashkin et al., 2019) is a dialogue-level emotion-tagged corpus, considering two participants as a *speaker* and a *listener*, and tagged with the speaker's emotion and its context.

In EmpatheticDialogues, not each utterance but each dialogue is annotated, which is not suitable for recognizing emotional transition throughout a dialogue. For Japanese, there is a Japanese version of EmpatheticDialogues called JEmpatheticDialogues (Sugiyama et al., 2021), which suffers from the same problem. In this work, we conduct utterance-level annotation like DailyDialog

[1] We will release our corpus and code at https://github.com/nlp-waseda/expr-exper-emo.

Length	# Dialogues	# Utterances
2	1,330	2,660
3	1,071	3,213
4	509	2,036
5	310	1,550
6	225	1,350
7	158	1,106
8	134	1,072
9	91	819
2-9	3,828	13,806

Table 1: The statistics of dialogues and utterances.

Label	Expressed		Experienced	
	Strong	Weak	Strong	Weak
Anger	430	1,349	124	870
Anticipation	1,906	4,229	1,215	4,068
Joy	1,629	3,672	1.553	4,549
Trust	247	1,732	520	3,455
Fear	252	942	123	846
Surprise	602	2,018	434	2,798
Sadness	1,227	2,936	889	3,037
Disgust	476	1,979	186	1,535
Any	6,371	12,215	4,705	12,515

Table 2: The statistics of utterances for each emotion label.

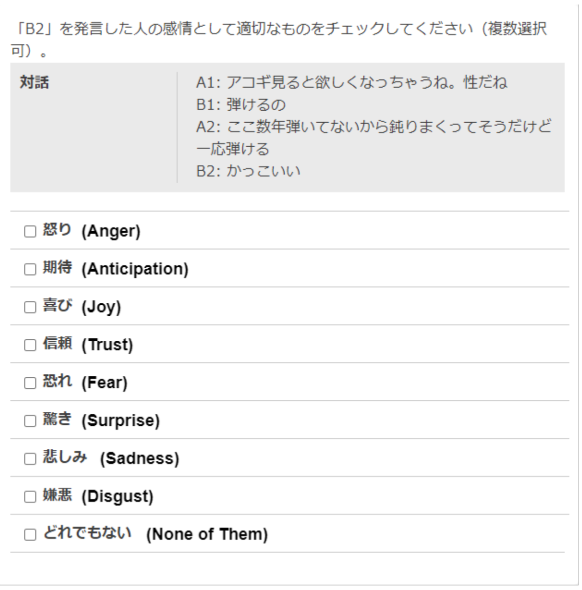
and EmotionLines. Although these corpora contain only the speaker's emotion (expressed emotion), we also annotate an utterance with the emotion of a person who hears it (experienced emotion). Furthermore, while an utterance has only one emotion label in these corpora, we allow multiple emotion labels to be tagged per utterance and also consider their strength.

There are also some studies toward developing emotion-aware dialogue systems. Smith et al. (2020) propose three skills for a human-like dialogue system: recognizing emotions, using knowledge (Dinan et al., 2019), and considering personality (Zhang et al., 2018). Furthermore, Roller et al. (2021) build a dialogue system capable of blending these three skills.

3 Corpus Building

3.1 Dialogue Collection

We collect dialogue texts from Twitter by considering the interaction between tweets and their replies by two users as a dialogue. To improve the text quality, we exclude tweets that contain images or hashtags and set the maximum number of utterances included in a dialogue to nine. We also apply several filters: excluding dialogues that contain special symbols, emojis, repeated characters, and utterances that are too short. Note that the reason why we exclude emojis is that they are relatively explicit emotional factors, and we intend to analyze emotions implied from usual textual expressions.



「B2」を発言した人の感情として適切なものをチェックしてください（複数選択可）。

対話	A1: アコギ見ると欲しくなっちゃうね。性だね B1: 弾けるの A2: ここ数年弾いてないから鈍りまくってそうだけど一応弾ける B2: かっこいい

- □ 怒り (Anger)
- □ 期待 (Anticipation)
- □ 喜び (Joy)
- □ 信頼 (Trust)
- □ 恐れ (Fear)
- □ 驚き (Surprise)
- □ 悲しみ (Sadness)
- □ 嫌悪 (Disgust)
- □ どれでもない (None of Them)

Figure 2: An example of the crowdsourced task. Checkboxes allow crowdworkers to select multiple emotions for an utterance.

We collected Japanese dialogues using this method. The numbers of dialogues and utterances are shown in Table 1. We obtained 3,828 dialogues that correspond to 13,806 utterances in total. Regarding the length of dialogues, the number of dialogues tends to decrease as that of utterances per dialogue increases.

3.2 Emotion Annotation

We adopt Plutchik's wheel of emotions (Plutchik, 1980) as annotation labels.[2] Specifically, our annotation labels consist of eight emotions: anger, anticipation, joy, trust, fear, surprise, sadness, and disgust. We annotate each utterance with two emotion categories: an expressed emotion, which is expressed by a speaker of the utterance, and an experienced emotion, which is experienced by a listener of the utterance. In other words, an utter-

[2] Ekman's six emotions (Ekman, 1992) and Plutchik's wheel of emotions (Plutchik, 1980) are commonly used in emotion-tagged corpora. Preliminary experiments by crowdsourcing showed that the latter is more appropriate for our crawled dialogues. In this work, therefore, we use eight emotions by Plutchik (1980).

Utterance	Expressed	Experienced
A1: 来週、車で京都行く 普通に観光してきます (I'm driving to Kyoto next week. We'll just do some sightseeing.)	{**Anticipation**, Joy}	{**Anticipation**}
B1: いいなぁ、京都 (I like Kyoto.)	{Anticipation}	{Anticipation, **Joy**}
A2: 楽しんできます (I'll enjoy it.)	{**Anticipation**, **Joy**}	{Anticipation, **Joy**}
B2: 行ったことないから純粋に羨ましい (I've never been there, so I'm genuinely jealous.)	{Anticipation}	{**Anticipation**, **Joy**}
A3: ないんや意外 (I'm surprised you haven't.)	{**Surprise**}	{Joy, Surprise}

Table 3: An example dialogue annotated with expressed and experienced emotions by crowdsourcing. The labels in bold indicate strong emotions.

Label	Expressed	Experienced
Anger	糞, せる, マジだ (shit, force, serious)	糞, うるさい, 居る (shit, noisy, exist)
Anticipation	教える, 願う, 待つ (teach, hope, wait)	待つ, 楽しみだ, 強い (wait, looking forward to, strong)
Joy	楽しい, 嬉しい, おもろい (joyful, glad, funny)	楽しい, 嬉しい, おもろい (joyful, glad, funny)
Trust	全然, 大丈夫だ, ちゃんと (at all, all right, properly)	やすみ, 教える, 大事だ (rest, teach, important)
Fear	怖い, やばい, どう (afraid, serious, how)	怖い, やばい, 危険だ (afraid, serious, dangerous)
Surprise	やばい, なんで, ？ (serious, why, ?)	居る, ビックリ, 年 (exist, surprise, year)
Sadness	泣く, 痛い, 悲しい (cry, hurt, sad)	泣く, 辛い, 痛い (cry, hard, hurt)
Disgust	悪い, 嫌いだ, 嫌だ (bad, hate, dislike)	悪い, 気持ち, 嫌だ (bad, surprise, dislike)

Table 4: Top-3 frequent words for each emotion label. An IDF filtering is applied to exclude common words.

ance is annotated with both subjective and objective emotions, which is similar to EmoBank (Buechel and Hahn, 2017) for non-dialogue texts. By annotating expressed and experienced emotions, we can trace the changes in the emotion surrounding both an utterance and a participant in a dialogue.

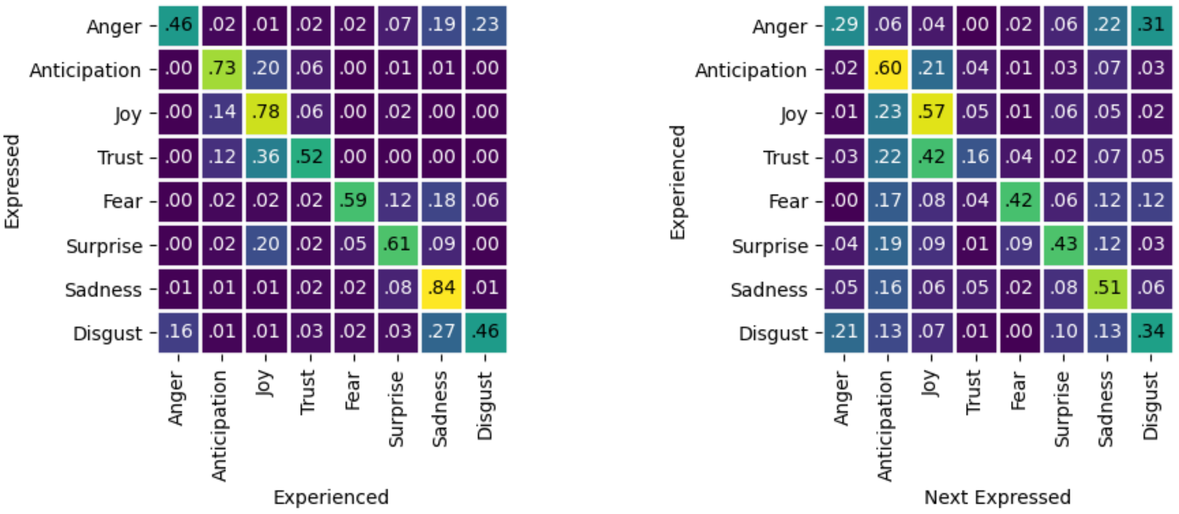
As a crowdsourcing platform, we use Yahoo! Crowdsourcing.[3] By showing the target utterance and its context, we ask seven workers whether the target utterance has a specified emotion or not about each emotion label for expressed and experienced emotion categories. For the expressed emotions, we ask which emotion a speaker expressed when saying the utterance. For the experienced emotions, we ask which emotion a listener experienced when hearing the utterance. Workers are allowed to select multiple emotion labels or none of them. An interface of the crowdsourcing task for expressed emotions is shown in Figure 2.

Because a view of expressed and experienced emotions can vary among annotators, we employ many workers per an utterance and aggregate their votes to obtain highly reliable annotations.[4] We consider strength for each emotion according to the number of workers' votes; emotions selected by more than half of the workers are regarded as strong, and ones selected by more than a quarter are regarded as weak. Note that the set of strong emotions is a subset of the set of weak ones. We expect that providing the emotions with intensity enables us to handle their granularity.

We applied the above emotion annotation method to our dialogue corpus. The number of utterances for each emotion is shown in Table 2. For the expressed emotion, 46.15% and 88.48% of the utterances are tagged with at least one strong and weak emotion, respectively. For the experienced emotion, the percentages are 34.08% and 90.65%, respectively. Approximately 90% of the utterances are accompanied by one or more emotion labels, and thus our corpus is consequently suitable for recognizing emotions in dialogues and analyzing their changes. In contrast to ours, for example, less than 20% of utterances are tagged with a specific emotion in DailyDialog (Li et al., 2017). Hence it is difficult to analyze emotion changes using such corpora with a small amount of emotion annotation. In addition, we can see a bias among the emotion labels for both expressed and experienced emotions, with more instances of anticipation and joy and fewer instances of trust and fear. An example of a dialogue with the annotation is shown in Table 3.

[3] `https://crowdsourcing.yahoo.co.jp/`

[4] For the expressed emotion, we ask workers to annotate the emotion of the speaker of an utterance. This annotation, however, is not strictly what a speaker had in mind but what the workers think a speaker would want to express, which can be considered *objective* in some sense. Having truly subjective annotation as the expressed emotion, like Kajiwara et al. (2021), is our future work.



(a) Expressed and experienced emotions (for a certain utterance).

(b) Experienced and next expressed emotions (for a certain person).

Figure 3: The confusion matrices of the relationship between expressed and experienced emotions. In this analysis, we focus on only the strong labels. Note that the matrices' elements are normalized in the row direction.

4 Corpus Analysis

4.1 Frequent Words for Emotion Categories and Labels

To investigate the characteristics of utterances with different emotions, we count words for each strong emotion label in our corpus. In this analysis, we identify words by the Japanese morphological analyzer Juman++ (Tolmachev et al., 2018). To exclude common words likely to appear for all emotions, we apply an IDF filtering. Specifically, words with IDF less than half of the maximum are ignored.

Top-3 words appearing for strong emotion labels are shown in Table 4. The same words tend to appear in the two emotion categories for joy and sadness. In contrast, the frequent words in the two categories are different for anticipation, trust, and surprise.

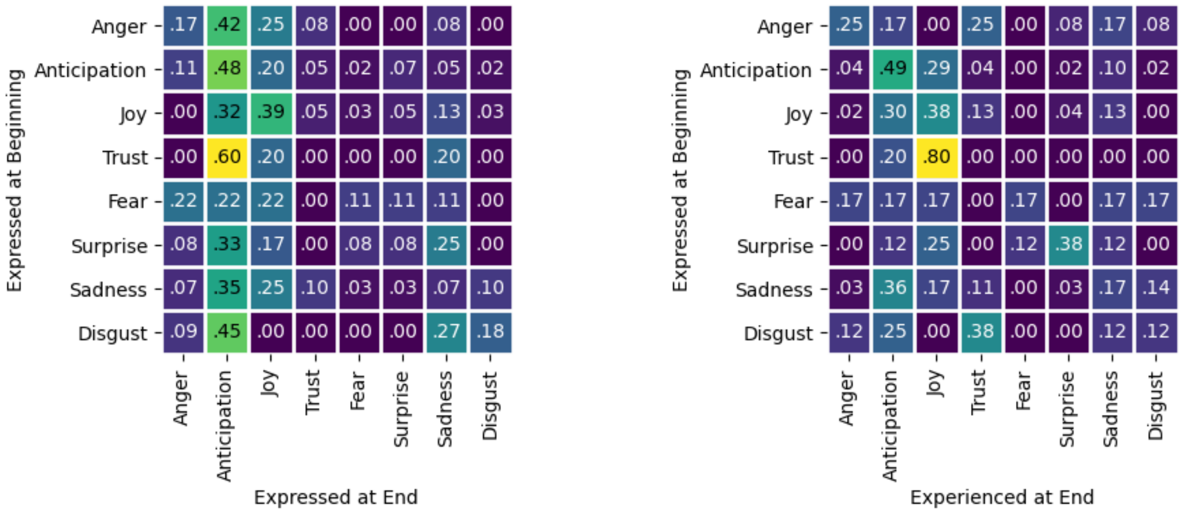
4.2 Relationship Between Expressed and Experienced Emotions

We annotated utterances with the expressed and experienced emotions. Here, we focus on the relationship between these two emotion categories. Specifically, we investigate the following two relationships:

a. The expressed emotion and the experienced emotion for the same utterance (different persons).

b. The experienced emotion for an utterance and the expressed emotion for the next utterance (the same person).

The confusion matrices for the strong emotion labels are shown in Figure 3, where the elements are normalized in the row direction. First, diagonal components of the two confusion matrices have large values, indicating that the same emotions are likely to occur both for the same utterance and for the same person. Figure 3a shows that people are likely to experience joy for an utterance of anticipation, trust, and surprise in addition to the same emotion. People also tend to experience disgust and sadness for anger and disgust, respectively. Figure 3b shows that after experiencing trust, people are more likely to express joy than trust. For an anger experience, people are more likely to express disgust than anger. Figures 3a and 3b reveal that the relationship of sadness is particularly different. For a certain utterance, sadness makes the other person feel sad in most cases, but for a certain person, anticipation in addition to sadness can be expressed after experiencing sadness. We speculate that when a person experiences sadness from the interlocutor, the person brings an utterance with anticipation to comfort them.



	Anger	Anticipation	Joy	Trust	Fear	Surprise	Sadness	Disgust
Anger	.17	.42	.25	.08	.00	.00	.08	.00
Anticipation	.11	.48	.20	.05	.02	.07	.05	.02
Joy	.00	.32	.39	.05	.03	.05	.13	.03
Trust	.00	.60	.20	.00	.00	.00	.20	.00
Fear	.22	.22	.22	.00	.11	.11	.11	.00
Surprise	.08	.33	.17	.00	.08	.08	.25	.00
Sadness	.07	.35	.25	.10	.03	.03	.07	.10
Disgust	.09	.45	.00	.00	.00	.00	.27	.18

(a) Expressed emotions at the beginning and end of dialogue.

	Anger	Anticipation	Joy	Trust	Fear	Surprise	Sadness	Disgust
Anger	.25	.17	.00	.25	.00	.08	.17	.08
Anticipation	.04	.49	.29	.04	.00	.02	.10	.02
Joy	.02	.30	.38	.13	.00	.04	.13	.00
Trust	.00	.20	.80	.00	.00	.00	.00	.00
Fear	.17	.17	.17	.00	.17	.00	.17	.17
Surprise	.00	.12	.25	.00	.12	.38	.12	.00
Sadness	.03	.36	.17	.11	.00	.03	.17	.14
Disgust	.12	.25	.00	.38	.00	.00	.12	.12

(b) Expressed emotions at the beginning and experienced emotions at the end of dialogue.

Figure 4: The confusion matrices for the emotion labels at the beginning and end of dialogue. In this analysis, we consider only the emotions of a person who begins a dialogue. Note that the targets are limited to the dialogues containing six to nine utterances, and the elements are normalized in the row direction.

4.3 Emotions at the Beginning and End of a Dialogue

To analyze the emotion changes through a dialogue, we compare emotions at the beginning and end of a dialogue. In other words, we see how the emotions of a person who starts the dialogue change through the dialogue. In this analysis, we focus on the following two relationships:

a. The emotion expressed first and the emotion *expressed* last by the same person.

b. The emotion expressed first and the emotion *experienced* last by the same person.

The confusion matrices for the strong emotion labels are shown in Figure 4. The targets are limited to dialogues containing six to nine utterances to analyze the emotion changes in long dialogues. Figure 4a shows that a speaker of the first utterance is likely to finally express anticipation and joy regardless of the first emotion. A speaker who first expresses surprise can express sadness through the dialogue. Figure 4b also shows that the first speaker can experience anticipation at the end of a dialogue. A person who first expresses anger and disgust tends to finally experience trust. From these two figures, we can see that a dialogue causes a person who first expresses fear to finally feel either a positive or negative emotion.

5 Experiments

5.1 Model Setup

We conduct experiments on expressed and experienced emotion recognition using our corpus. We solve a regression task of each emotion intensity for an utterance with its context for the emotion recognition task. We assign 0, 1, and 2 for none, weak, and strong emotion labels, respectively, and let a model regress these values for each emotion. As such, we train two separate models for expressed and experienced emotions with the mean squared error loss:

$$\mathcal{L} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} (y_{ij} - t_{ij})^2, \qquad (1)$$

where N is the number of samples and K is the number of emotion labels. y_{ij} is the output from the model for the jth label of the ith sample, and t_{ij} is its gold label.

We adopt a Japanese pre-trained BERT model and fine-tune it. We compare two pre-trained models from Kyoto University[5] and one from NICT[6]. We use the WWM and BPE versions for Kyoto University's and NICT's BERT models, respectively.

[5] https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese

[6] https://alaginrc.nict.go.jp/nict-bert/index.html

Model	Expressed	Experienced
Kyoto (base)	58.84/44.33	53.60/41.84
Kyoto (large)	60.85/45.16	55.09/42.94
NICT	**61.50/46.05**	**56.23/43.88**

Table 5: The results of regression for expressed and experienced emotions. The metrics are Pearson's and Spearman's correlation coefficients.

Label	Expressed	Experienced
Anger	50.21/33.80	38.11/23.80
Anticipation	62.76/**55.55**	57.46/51.22
Joy	**67.25**/55.22	**61.92/54.47**
Trust	41.15/36.69	43.91/40.48
Fear	59.09/31.47	49.60/24.90
Surprise	49.86/39.58	40.58/33.86
Sadness	63.70/51.50	55.48/43.88
Disgust	47.76/38.18	37.32/28.13

Table 6: The correlation coefficients for each emotion label. The metrics are Pearson's and Spearman's correlation coefficients. The scores are from the NICT model that achieved the highest performance in Table 5.

Input utterances are segmented into words with Juman++ (Tolmachev et al., 2018) and tokenized into subwords by applying BPE. We join utterances with `[SEP]` and append `[CLS]` and `[SEP]` to the beginning and end, respectively. As there are two participants in a dialogue, we give each utterance a segment ID of 0 or 1. It provides the models with the information about the speaker of an utterance. Based on a series of utterances joined with `[SEP]`, we predict an emotion label for the last utterance. The vector corresponding to `[CLS]` is passed to a fully-connected layer, and an eight-dimensional vector representing the eight emotions is obtained. Each of the elements is supposed to regress the intensity of each emotion.

Since we are dealing with a regression task, Pearson's and Spearman's correlation coefficients are used as evaluation metrics. The dialogues in our corpus are split into 8:1:1, corresponding to training, validation, and test sets. We fine-tune our models for three epochs and evaluate them on the test set. The implementation of the models is based on HuggingFace Transformers[7]. The models are trained using NVIDIA Tesla V100 SXM2 GPU.

5.2 Results

For the regression task defined in Section 5.1, the correlation coefficients for each model are shown in Table 5. In terms of performance, the NICT model achieved the best score across all values. For the values regarding expressed and experienced emotions, the performance of the experienced emotion is inferior to that of expressed emotion in all models. This result indicated that it is more difficult to recognize the experienced emotion than the expressed emotion.

The correlation coefficients for each emotion inferred by the NICT model are shown in Table 6. For both the expressed and experienced emotions, the highest scores were achieved for *anticipation* and *joy*. In contrast, the emotions with lower values were *trust* and *fear* for the expressed emotion and *anger* and *disgust* for the experienced emotion. From Tables 6 and 2, we can see that the larger the number of the samples for an emotion is, the higher the correlation coefficient becomes. As a case study, we show example dialogues and their emotions predicted by the NICT model in Table 7.

5.3 Multi-Task Learning

Our analysis in Section 4.2 indicated that there is a correlation between expressed and experienced emotions. Therefore, we consider training a single model for recognizing both the emotion categories. The information for solving the two similar tasks is expected to allow a model to improve the performance of each other (Liu et al., 2019). We provide a model with two separate fully-connected layers for the tasks and train them simultaneously, where the inputs are the same as those in Section 5.1. Here, the mean of the losses for expressed and experienced emotions is optimized:

$$\mathcal{L}_{\text{multi-task}} = \frac{\mathcal{L}_{\text{expressed}} + \mathcal{L}_{\text{experienced}}}{2}. \quad (2)$$

Based on Figures 3a and 3b, we consider multi-task learning of expressed and experienced emotions for a certain utterance and a certain person. For the relationship in a certain person, we use the experienced emotion of an utterance and the expressed emotion of the following utterance. We also conduct experiments on the cases where the training and test sets are different from each other. In such a case, for example, expressed emotions are used

[7] https://huggingface.co/transformers/

Dialogue	Predicted	Gold
A1: ゲームの検証してる人が検証してほしいことあれば言ってください的なこと言ってたから依頼したら無視されて悲しくなったのはいい思い出 (I have a good memory of a guy who was verifying a game and said if there was anything he wanted verified, please let him know, so I made a request and he ignored it, which made me sad.) B1: それは悲しいね (That's sad.)	**Strong sadness**	**Strong sadness**
A1: youtubeでバーのマスターが氷砕いてる動画見てボーッとしてる (I've been watching videos of bar masters crushing ice on youtube and I'm in a daze.) B1: なんかしてよ (Do something.) A2: そのうちこういうときにツイキャスをしようかなと思っておる (One of these days I'm going to do a tweak for this.) B2: 天才の発想 スマホでも見やすいから助かる (It's a genius idea, and it's easy to watch on my phone.)	Weak anticipation and **joy**	Strong **joy** and weak trust
A1: 今、部活終わって帰るとこやけど 雨やばいしかっぱ持ってきてないし 最悪 (I'm on my way home after club activities, but it's raining and I didn't bring my hat, so that sucks.) B1: わたしも学校出た瞬間大雨降ってきた (I'm going back to school now, but it's raining really hard and I didn't bring my jacket.)	Strong surprise	Strong sadness

Table 7: Example dialogues with predicted and gold expressed emotions. The predicted emotion labels are taken from the predictions of the NICT model, which predicted an emotion label for the last utterance of each dialogue.

Train\Test	Expressed	Experienced
Expressed	61.50/46.05	52.89/40.91
Experienced	55.49/43.34	56.23/43.88
Multi-Task	**62.20/46.63**	**57.35/45.01**

Table 8: The results of multi-task learning with expressed and experienced emotions. The metrics are Pearson's and Spearman's correlation coefficients.

Train\Test	Experienced	Next Expressed
Experienced	54.62/43.47	29.53/25.46
Next Expressed	43.32/35.27	33.91/28.31
Multi-Task	**55.75/49.50**	**35.17/30.49**

Table 9: The results of multi-task learning with experienced and next expressed emotions. The metrics are Pearson's and Spearman's correlation coefficients.

for training, but experienced emotions are used for testing.

The correlation coefficients for an utterance and a speaker by multi-task learning are shown in Tables 8 and 9, respectively. First, the scores when the training and test sets are different from each other are lower than those when they are the same. This gap indicates the significance of annotating utterances with expressed and experienced emotions separately. In all columns, the multi-task models achieved higher performance than the single-task models. Especially, in Table 9, the multi-task scores for both the two tasks are higher than the single-task baselines by one point. In other words, expressed, experienced, and next expressed emotions have the information for helping the recognition of each other.

6 Conclusion

We proposed a method to build an emotion-tagged multi-turn dialogue corpus to help machines recognize emotional transition in a dialogue. Dialogues between two speakers are collected from Twitter, and each utterance is annotated with emotions by crowdsourcing. In the annotation process, we consider the emotions expressed by a speaker who said the utterance and the emotion experienced by a listener who heard the utterance. In addition, the labels are provided with their intensity, representing the granularity of emotions.

We built a Japanese emotion-tagged dialogue corpus and analyzed it. The results showed the characteristics of words for each emotion, the correlation between the emotions about a certain utterance and speaker, and the tendency for speakers to become positive through a dialogue. We also developed emotion recognition models for expressed and experienced emotions based on the Japanese pretrained BERT models. The experimental results indicated that it is more difficult to recognize a listener's emotion than a speaker's emotion. Multi-task learning of expressed and experienced emotions improved the performance of the two emotion recognition tasks about an utterance and a speaker.

For our future work, we will tackle response generation based on predicted emotions. With our corpus, a dialogue system is expected to predict which emotion it experiences from a given utterance and which emotion it should express for the next utterance. Once such emotions are recognized, the dialogue system should be able to generate an appropriate response depending on the predicted expressed emotion.

The corpus in this work is annotated only with expressed and experienced emotions about an utterance. In addition to the emotion annotation, we should also consider dialogue situations (Rashkin et al., 2019). The cause of a dialogue or an utterance helps recognize a speaker's emotion and how it changes. We can also consider non-emotional annotation, such as a dialogue's topic and an utterance's intention (Li et al., 2017). The relationship between emotions and non-emotional factors is also important for machines to better recognize a speaker's emotion.

Acknowledgements

This work was supported by a joint research grant from LINE Corporation.

Ethical Considerations

We built the dataset by collecting texts from Twitter and annotating them by crowdsourcing. For crowdsourcing, we employed 3,847 workers. It took approximately five minutes for a task of annotating 10 utterances. Every worker was paid 4 JPY per 10 utterances, and in total, the built dataset costs 195,700 JPY. Since the dataset was collected from Twitter, it may include contents that are harmful for some of the dataset or its application users. For building the dataset through the Twitter API and crowdsourcing, we did not include any sensitive information that allows personal identification.

The dataset or models trained on it enable downstream applications to infer the emotions of their users, resulting in facilitating communication between the users and the applications. In terms of dialogue systems, this ability is considered valuable for both task-oriented and non-task-oriented dialogue systems. For example, it assists the user in decision-making and solves the user's worry and trouble. In contrast to such benefit, it is difficult for the model to infer the emotion accurately, with the relatively small dataset. Therefore, prediction errors by the model, especially for sensitive utterances or negative emotions, may bring harmful experiences on the users.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing*, pages 152–165, Cham. Springer International Publishing.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Vicki Liu, Carmen Banea, and Rada Mihalcea. 2017. Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2019. Positive emotion elicitation in chat-based dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):866–877.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.