Investigating the relationship between dialogue and exchange-level impression

Wenqing Wei
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
wqwei@jaist.ac.jp

Sixia Li
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
lisixia@jaist.ac.jp

Shogo Okada
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
okada-s@jaist.ac.jp

ABSTRACT

Multimodal dialogue systems (MDS) have recently attracted increasing attention. The automatic evaluation of user impression with spoken dialog at the dialog level plays a central role in managing dialog systems. A user usually forms an overall impression through the experience of each exchange of turns in the conversation. Thus, the user's exchange-level sentiment should be considered when recognizing the user's overall impression of the dialog. Previous research has focused on modeling user impressions during individual exchanges or during the overall conversation. Thus, the relationship between user sentiment at the exchange level and user impression at the dialog level is still unclear, and appropriately utilizing this relationship in impression analysis remains unexplored. In this paper, we first investigate the relation between sentiment at the exchange level and 18 labels that indicate different aspects of the user impression at the dialog level. Then, we present a multitask learning model (MTL) that uses exchange-level annotations to recognize dialog-level labels. The experimental results demonstrate that our proposed model achieves better performance at the dialog level, outperforming the single-task model by a maximum of 15.7%.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

KEYWORDS

Dialogue-level impression, Exchange-level impression, multitask learning, deep learning neural networks, impression, sentiment, multimodel dialogue system

ACM Reference Format:
Wenqing Wei, Sixia Li, and Shogo Okada. 2022. Investigating the relationship between dialogue and exchange-level impression. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536221.3556602

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556602

1 INTRODUCTION

With the development of natural language processing (NLP) and speech recognition, dialog systems such as speech assistant systems, information guide systems in public locations, and intelligent customer service systems have begun to play crucial roles in our lives. An important step in improving the quality of dialog systems is to evaluate the user's impression of the system. Many previous studies have used objective evaluation metrics to assess dialog systems. Previous research on user impressions can be divided into two categories: dialog-level and exchange-level user impression modeling. Exchange-level methods are designed to evaluate the user's impression at any point in a particular conversation. The main purpose of exchange-level user impression evaluations is to track the user's internal impressions (such as their sentiment) and to adapt verbal and nonverbal responses according to the user impression. A good dialog system should be not only coherent and appropriate but also engaging [26]. The user's overall impression of the conversation system is also influenced by the experience of each exchange, and the user's self-sentiment at the exchange level may be helpful for determining the user's overall evaluation of the dialog.

Previous studies [11, 12, 21] have recognized user interactions at the dialog and exchange levels separately, and few studies [5] have explored and utilized the relationship between the two. Bodigutla et al. [5] used multiple tasks to recognize turn- and dialog-level impression ratings for a given task-oriented dialog proving that dialog-level labels are beneficial for evaluating user satisfaction at the exchange level. However, this study focused on modeling user satisfaction in text-to-text dialog systems rather than multimodal systems and did not explore the relationship between the dialog and exchange levels, which remains unclear. In particular, the type of user knowledge that is shared between dialog-level and exchange-level sentiment is important in developing multimodal dialog systems.

Based on this background, we use a publicly available multimodal dialog dataset that contains multimodal data [16], including audio, body, visual, and transcript data, as well as two types of user sentiment labels to evaluate user impressions of the system. The dataset was collected with a non-task-oriented dialogue system with the annotations of dialogue level and exchange level sentiments. Therefore, this dataset allows us to investigate the relationship between the dialogue level and exchange level. Based on the findings of the correlation analysis between the two types of labels, we present a multitask model for recognizing dialog-level annotations that includes two tasks: recognizing the self-sentiment at the exchange

level and recognizing the dialog-level label. To validate the effectiveness of the proposed multitask model, we compare the dialog-level labels of the single-task and multitask models. The comparative analysis in Section 6.2 indicates that considering user sentiment at the exchange level was helpful for recognizing the user impression at the dialog level. Furthermore, we compared the results of our proposed model with those of other works using the same database. We demonstrate that our multitask model achieved better performance in Section 7.2. The main contributions of this study can be summarized as follows:

Exploration of the relation between exchange-level labels and dialog-level labels: To explore the relationship between the exchange level and the dialog level, we use a dataset and annotate the user impressions at the exchange and dialog levels. We first analyze the relationship between user sentiment at the exchange and dialog levels in Section 3.3. Then, we investigate the effect of the correlation coefficient on the multitask performance in Section 7.2. The comparative analysis demonstrates that the correlation generally has a positive effect on the multitask performance.

Sequential multitask learning (MTL) of both exchange-level and dialog-level labels: Sequential MTL [1] enables the model to utilize information from various tasks to learn important common information between different tasks. This characteristic allows MTL to train the model to handle one task while accounting for other sub-tasks. However, multiple labels are assigned to the same dimensional data input in basic MTL. Thus, we need to handle multiple labels that are assigned to different units (exchange and dialog) in this study. To train the model using labels assigned to different units, we utilize the sequence modeling method. We use long short-term memory (LSTM) and gated recurrent units (GRUs) as baselines. We show the impact of MTL in Section 6.2, and the results show that the MTL strategy improved the recognition accuracy on almost all dialog-level tasks.

2 RELATED WORKS

With the development of social signal processing, human-computer interaction applications are increasingly used in people's daily lives. According to whether the dialog has a goal, the dialog system can be roughly divided into two categories: task-oriented dialog systems and open-domain non-task-oriented dialog systems. Task-oriented systems assist users in solving specific tasks as efficiently as possible [24]. In contrast, non-task-oriented systems do not have a specific task, and their main purpose is to entertain users with open-domain chats [14]. Non-task-oriented systems are used in many fields, such as assisting elderly people and those learning a second language. In a non-task-oriented dialog system, it is more important to engage the user in the interaction as long as possible and to ensure that the users returns as often as possible rather than to respond to the user correctly. Previous studies have shown that people spontaneously adjust their facial expressions, postures, pronunciation, and speech rates during conversation [6, 17, 18], which demonstrates that dialog should be able to capture the unspoken intentions, attitudes, and emotions of interlocutors, especially in non-task-oriented dialog systems. [25] proposed a multimodal non-task-oriented dialogue system that improved user experience by assessing the multimodal behavior of users.

The evaluation of the performance of a dialog system is one crucial component of managing dialog. Non-task-oriented dialog systems, such as an open-domain dialog systems, cannot set clear goals for the dialog; thus, it is more difficult to evaluate task accomplishments than in task-oriented dialog systems. To address this issue, recent research has focused on recognizing user-centered criteria, such as satisfaction and interaction quality annotated by users [22, 23]. In recent decades, many researchers have focused on evaluating user status. Most can be divided into two categories: exchange-level evaluations and dialog-level evaluations.

For exchange-level evaluation tasks, as the user's impressions can change dynamically during dialog exchanges, it is necessary to capture these dynamic changes in real time so that the system's next action adopts a dialog strategy based on the user's impressions. Schmitt and Hara et al. [11, 21] used support vector machines and n-grams to predict the quality of interactions in ongoing dialogs at the exchange level. Engelbrecht et al. [9] used hidden Markov models (HMMs), and the user's opinion was regarded as a continuously evolving process. Historical context plays a crucial role in conversations and is beneficial for recognizing user satisfaction and considering temporal features at different levels. To overcome the limitation of handcrafting temporal features, Ultes et al. [22] developed a recurrent neural network for recognition user satisfaction. To evaluate different aspects of the user's impressions, Hirano et al. [13] proposed a multitask deep learning neural network model that used multimodal features and deep neural networks (DNNs) to recognize the three exchange-level features: (1) the user's interest label, (2) the user's sentiment label, and (3) the topic continuance label toward the spoken dialog system.

The main purpose of the dialog level evaluation task is to learn dialog strategies to maximize the overall impression of the dialog, which is also helpful for identifying problematic conversation topics that led to user dissatisfaction. Higashinaka et al. [12] used overall dialog ratings to estimate dialog-level quality by using HMM and overcame the limitation by using task success [20] as dialog evaluation criteria. To estimate user satisfaction in conversations that span multiple domains, Bodigutla et al. [4] used new domain-independent feature sets (the aggregate topic popularity and the diversity of topics in a session) to estimate the user satisfaction at both the turn and dialog levels. Wei et al. [23] proposed a multimodal user satisfaction recognition model to evaluate non-task-oriented dialog systems at the dialog level by using automatic multimodal features. Furthermore, they investigated the contribution of different modalities to user satisfaction at the dialog level.

All the above works recognized user impression on dialog and exchange levels separately. To utilize the relationship between the two, this study proposes a multimodal model for recognizing the dialog level user impressions by considering the user's sentiment at the exchange level that is suitable for evaluating non-task-oriented dialog systems. We first explore the relationship between 18 dialog labels and the user's sentiment at the exchange level and then utilize MTL models, which enable the model to recognize self-sentiment while considering the user's overall impression at the dialog level. Fig. 1 shows an overview of this research.

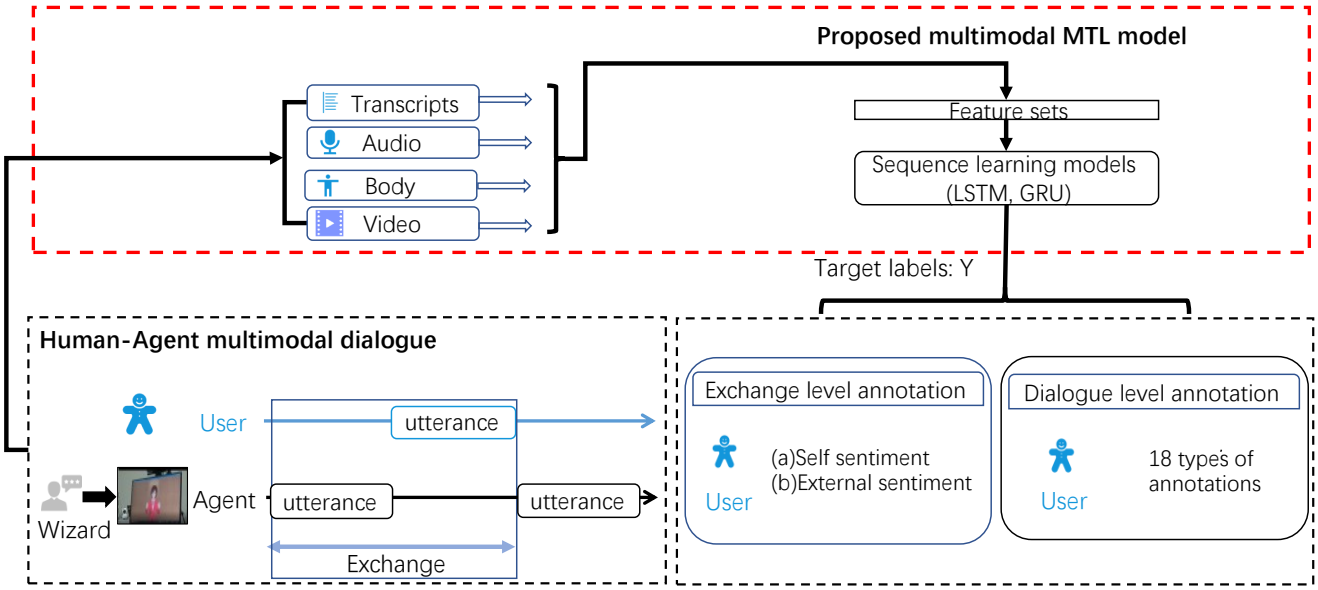

Figure 1: Overview of the MTL multimodal model for recognizing user impressions.

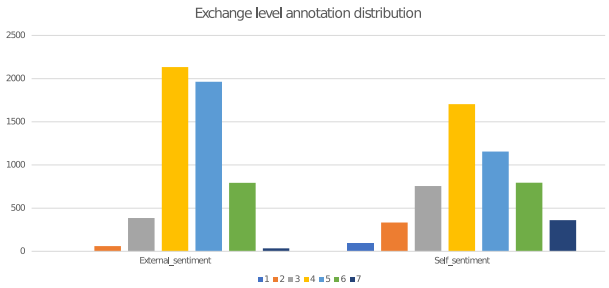

Figure 2: The rate distribution of the exchange-level annotations.

3 DATA DESCRIPTION

3.1 Data

Following [16, 23], the Hazumi1902 and Hazumi1911 data corpora were employed in this research. These two corpora include 60 participants (25 males/35 females aged 20-60 years). To reduce the effect of participants having different preferences on various topics, The behavior of the participants was recorded with a video camera and Microsoft Kinect V2 sensor.

3.2 Annotations

3.2.1 Dialog level annotations. The dataset used a questionnaire with 18 labels relating to the user's impression of the dialog, as proposed in [3]. The questionnaire measured cognition and rapport in interpersonal communication. The 18 items were "well-coordinated", "boring", "cooperative", "harmonious", "unsatisfying", "uncomfortably paced", "cold", "awkward", "engrossing", "unfocused", "involving", "intense", "friendly", "active", "positive", "dull", "worthwhile", and "slow". Each label was evaluated on an eight-point scale from 1 to 8 by the users after the dialogue.

3.2.2 Exchange level annotations. In this section, we describe the different exchange-level labels in detail.

External sentiment: In this work, an exchange was defined as the part that begins at the start time of a system utterance and ends at the start time of the next system utterance. Human coders annotated the external sentiment according to the participant performance during each exchange with a score ranging from 1 (participant seems bored with the dialog) to 7 (participants seem to enjoy the dialog), and five experts annotated the external sentiment labels. The distribution of the self-sentiment and external sentiment (average score annotated by the five experts) on the exchange-level labels are shown in Fig. 2.

Self-sentiment: This annotation was similar to the external sentiment annotation. Self-sentiment labels were assigned as scores ranging from 1 (want to stop talking, confused about the systems utterances) to 7 (enjoy talking, satisfied with the talk) and were annotated by the participants themselves.

3.3 Data analysis

In this study, we focus on exploring the relationship between the overall exchange sentiment within a dialog and the dialog-level sentiment. This relationship can reflect whether certain exchange sentiments lead to certain dialog sentiments from the viewpoint of the whole dialog. We computed the Pearson correlation coefficients between the average exchange labels (values) over the time-series exchanges and the dialog labels of all dialogues and investigated the distribution of the coefficient values.

Moreover, as seen from the above definitions of dialog-level labels, the dialog-level labels describe the user impression with both positive and negative annotations. Since different polarity

Table 1: Pearson correlation coefficient and P-value (p) results between exchange-level sentiments and dialogue-level sentiments. (a) shows the coefficients between exchange-level sentiments and positive dialogue-level annotation; (b) shows the Pearson correlation coefficients between exchange-level sentiments and negative dialogue-level sentiments. (represents $p < 0.001$, * represents $0.001 < p < 0.05$; if $p > 0.05$, no symbol is shown.)**

(a) Pearson correlation coefficients of positive dialogue-level annotations

	Third-party sentiment	Self-sentiment
Well-coordinated	+0.222	+0.297 *
Cooperative	-0.002	+0.110
Harmonious	+0.058	+0.309 *
Engrossing	+0.179	+0.359 *
Involving	+0.154	+0.298 *
Friendly	+0.042	+0.283 *
Active	+0.068	+0.253 *
Positive	+0.154	+0.245
Worthwhile	+0.037	+0.531 **
Average	+0.101	+0.299

(b) Pearson correlation coefficients of negative dialogue-level annotations

	third-party sentiment	Self-sentiment
Boring	-0.135	-0.448 **
Unsatisfying	-0.123	-0.289 *
Uncomfortably paced	+0.049	-0.279 *
Cold	-0.270 *	-0.528 **
Awkward	-0.156	-0.362 *
Unfocused	-0.215	-0.261 *
Intense	-0.237	-0.164
Dull	-0.263	-0.453 **
Slow	-0.190 *	-0.386 *
Average	-0.171	-0.352

labels describe opposing annotations, we divide the dialog-level labels into two categories to precisely investigate the different annotation polarities. The positive category includes the labels well-coordinated, cooperative, harmonious, engrossing, involving, friendly, active, positive, and worthwhile. The negative category includes the labels boring, unsatisfying, uncomfortably paced, cold, awkward, unfocused, intense, dull, and slow.

Table 1 (a) lists the coefficient value between exchange-level sentiments and positive dialog-level annotations, and Table 1 (b) lists the coefficient value between exchange-level sentiments and negative dialog-level annotations. Each row indicates a dialog-level annotation, and each column indicates an exchange-level sentiment. The intersection between a row and a column represents the coefficient value between an overall exchange-level sentiment and a dialog-level annotation. The average shows the average coefficient value of the coefficients of a given polarity corresponding to a given exchange-level sentiment.

As seen in the table, all coefficients between the exchange-level sentiment and the positive dialog-level annotations are positive, except the coefficient between the third-party sentiment and the cooperative label, which is negative but close to zero. The average coefficient between the third-party sentiment and the positive dialog-level annotations is 0.101. The average coefficient between the self-sentiment and the positive dialog-level annotations is 0.299. On the other hand, all coefficients between the exchange-level

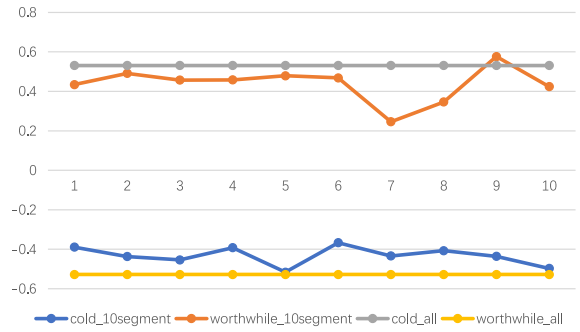

Figure 3: The Pearson correlation coefficient between each segment and dialog-level label (worthwhile and cold).

sentiment and the negative dialog-level annotations are negative. The average coefficient between the third-party sentiment and the negative dialog-level annotations is -0.171. The average coefficient between the self-sentiment and the negative dialog-level annotations is -0.352. These results demonstrate that dialog-level annotations are closely related to exchange-level sentiments. A higher overall exchange-level sentiment leads to a positive dialog-level annotation, while a lower overall exchange-level sentiment leads to a negative dialog-level annotation. Almost all exchange-level self-sentiment and dialog-level annotation pairs have a significant correlation, with $p<0.05$; however, most third-party sentiment pairs are not significant ($p>0.05$). This result indicates that exchange-level self-sentiment labels correlate more and have more common information with dialog-level labels. Moreover, compared with the self-sentiment label annotated by the user themselves, the external sentiment annotated by the five experts was more time-consuming and expensive. Thus, our study used the self-sentiment as a sub-task target label.

Table 1 shows that the "worthwhile" label obtained the highest correlation with the self-sentiment among all positive annotations, while the "cold" label had the highest correlation with the self-sentiment label among all negative annotations. To explore the correlation between different conversation segments and dialog-level annotations, we divided the dialog into 10 segments and used Pearson correlations to compute the average coefficient value of each segment. Fig. 3 shows the correlation coefficient of the cold and worthwhile labels for each segment and the overall conversation with the self-sentiment label. We observed that different segments had distinct relations at the dialog label, and few segments (worthwhile-10th) had higher correlations at the dialog level. Compared with the correlation between the average value of each conversation segment and the self-sentiment, the average value of the overall dialog had a higher correlation with dialog-level annotations, which indicated that considering all exchanges is better for recognizing the dialog label. For this reason, all exchange-level information was utilized in all experiments.

We confirm that the dialog-level annotations are closely correlated with the exchange-level self-sentiment and suggest that considering the exchange-level sentiment can improve dialog-level

annotation recognition. For this reason, in this study, we recognize user impressions by considering user self-sentiment at the exchange level.

4 METHODS

4.1 Feature extraction

4.1.1 Audio features. For the acoustic modality, we use the speech feature extractor OpenSMILE [10] to extract acoustic features at the exchange level. The acoustic features correspond to the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which achieves high performance in emotion-related fields. The features of each speaker are extracted and normalized. Because these acoustic features represent the performance of an entire exchange, we apply the same acoustic features to utterances that have different labels in one exchange.

4.1.2 Linguistic features. For the linguistic modality, we extracted two types of linguistic features from transcriptions of spoken dialog contents.

Part of speech: The sentences were segmented into words and annotated with universal part-of-speech (POS) tags by using Stanza NLP [1]. The number of different POS tags for each sentence are counted. We use a 17-dimensional vector as a sparse representation of 17 POS tags.

Bidirectional Encoder Representations from Transformers (BERT) [8]: Language model pretraining has been proven to be useful for learning universal language representations. A model pretrained on Japanese text (using Wikipedia) [15] was employed in this work. We use this model to extract features from text at the exchange level, yielding a 768-dimensional text representation vector. Thus, we obtain a 785-dimensional linguistic feature vector.

4.1.3 Body features. For body features, this work uses three-dimensional coordinates of each joint in the upper body, which were estimated with a Microsoft Kinect v2 sensor. Five points of body motion are employed: the left shoulder, right shoulder, left hand, right hand, and head. We denote the three-dimensional coordinate of each body point in frame t as $w(t) = x, y, z$ and the time between frames as t_1. We calculate the absolute value of the velocity between two frames as $|v(t)| = |w(t+1) - w(t)|$ and the absolute value of the acceleration between two frames as $|a(t)| = |v(t) - v(t-1)|$. We calculate the velocity and acceleration to coordinate the data of the 5 body points in all frames. After $v(t)$ and $a(t)$ are calculated, we use the maximum value of the acceleration and the maximum, mean, and standard deviation of the velocity in each exchange turn as body activity features. Thus, the body activity feature set has a total of 20 dimensions.

4.1.4 Visual features. For extracting visual features, we used OpenFace [2] software.

Facial landmark features: OpenFace outputs three-dimensional coordinates of 68 facial landmarks in each frame. In this study, ten facial landmarks were selected: two on each eye, four on the mouth, and two on the eyebrow. We adopted the same method as used for tracking body features. The maximum acceleration value and the maximum, mean, and standard deviation of the velocity were extracted for each user exchange and used as visual features. Thus, we obtained a 40-dimensional vector.

Action units: Facial expressions display emotional states that objectively describe facial muscle activation [7]. To extract the facial expression, this study used OpenFace software to obtain 18 action units (AUs) that were rated between 0 and 1, which indicate absence and presence, respectively. Then, we calculated the average of each AU during the exchange to obtain the facial AU features (18-dimensional). Overall, 58 dimensions of visual features were used in this study.

[1]https://github.com/stanfordnlp/stanza

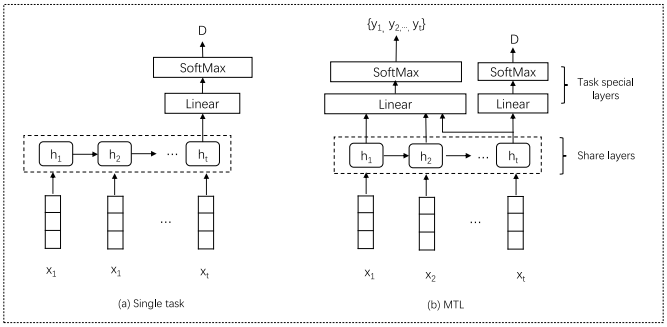

Figure 4: The structures of the signal task model and multitask model.

4.2 Model

4.2.1 Single task deep learning neural network (baseline). User multimodal behavior dynamically changes during a conversation. To preserve the sequential information, we use the LSTM and GRU methods to recognize the user impression. As described in Section 4.1, different unimodal features (audio a_t: 88-dim., linguistic l_t: 785-dim., body b_t 20-dim. and video v_t: 58-dim.) were extracted from the t-th exchange. We use the early fusion method to concatenate different unimodal features, generating the exchange-level multimodal feature $x_t = [a_t, v_t, b_t, v_t]$. The multimodal feature X = (x_1, x_2, ..., x_t) was used as the input of these neural network models. In all models with one recurrent layer and 128 units, we obtained a 128-dimensional hidden state from the recurrent layer. The recurrent layer was followed by a fully connected layer, which projected the output (128-dimensional). At the end of the model output layer, which contained two units, the log- softmax function was used to output the probabilities of different user impressions. As shown in Fig. 4 (a), the output at the final moment h_t can be regarded as a representation of the whole sequence, which uses a fully connected layer followed by a softmax nonlinear layer to predict the probability distribution over different classes.

4.2.2 Multitask deep learning neural network (proposed model). MTL is a machine learning approach that simultaneously solves multiple learning tasks by exploiting commonalities and differences across tasks [1]. An advantage of the multitask model is that it utilizes correlations among dialog-level and exchange-level tasks, improving the classification performance by learning several tasks in parallel. The key factor of MTL is the sharing scheme in the latent feature space. In a neural network-based model, the latent features

can be regarded as the states of the hidden neurons. In general, MTL models are composed of two parts: shared layers and task-specific layers. The lower layers are shared across all tasks, and there are several task-specific layers. In the multitask model, we use a single recurrent layer with 128 units as the shared layers. These layers extract features at both the exchange level and dialog level for the tasks shown in Fig. 4 (b). In the task-specific layers, for the exchange sentiment task, we obtained 128-dimensional hidden states H = (h_1, h_2, ..., h_t) from the recurrent layer, which was followed by a fully connected layer that projected the output (128- dimensional). The output layer contains two units, and the log-softmax function of each hidden unit outputs the exchange-level task-specific layer at time step t, which are the probabilities for different exchange self-sentiments x_t. For the dialog-level user impression recognition task, the structure is the same as in the single task, and the mathematical formula of the model can be described as follows:

$$Share\ layer : h_t = LSTM(x_t W_e, h_{t-1}) \tag{1}$$

$$Exchange\ level\ task\ special\ layer : y_t = Softmax(h_t W_y + b_y) \tag{2}$$

$$Dialogue\ level\ task\ special\ layer : D = Softmax(h_t W_D + b_D) \tag{3}$$

Equation 4 shows the multitask loss function of the multitask model. L_e and L_d are the mean square error losses computed for the exchange-level and dialog-level label ratings, respectively. The values λ and (1-λ) are interpreted as the loss weights of the dialog-level task and exchange-level task, which were set manually.

$$L = \lambda * L_d + (1 - \lambda) * L_e \tag{4}$$

5 EXPERIMENT

The dialog-level user impression recognition task and exchange-level sentiment recognition task are both time-series tasks. Thus, considering the time-series information is beneficial for improving model performance. Recurrent neural networks are mainly used for tasks that involve sequential inputs, such as time-series predictions. This work uses LSTM and GRUs as baselines to model the sequence of multimodal behaviors. To eliminate the influence of unbalanced data, we adopt 5-fold cross-validation to train the models, yielding 5 groups of evaluation results. The mean of the 5 groups was computed and used as the final result, and the F1-score of the label weights was used as the evaluation metric. According to previous works, linguistic features are key descriptors in recognizing user satisfaction. In this work, we used a unimodal model with a linguistic feature set as the baseline model. We compare the accuracy with the following 5 feature sets to analyze the contribution of each modality to the recognition of the dialog-level label:

(1) L: model trained with linguistic features (baseline)
(2) L+A: model trained with linguistic features + acoustic features
(3) L+B: model trained with linguistic features + body features
(4) L+V: model trained with linguistic features + visual features
(5) ALL: model trained with acoustic features + body features + visual features + linguistic features

5.1 Experimental settings

The experiment was performed for the different modalities based on the (1) dialog-level labels and (2) exchange-level user self-sentiment labels. The binary classification datasets were developed as follows. The dialog-level label annotated scores (1-8) were converted to binary values (high and low) with a threshold of 4 (neutral state). The self-sentiment, which is rated between 1 and 7, was converted to binary values (high/low) with a threshold of 4. The number of high/low point for self-sentiment label on exchange level were 2882/2311.

Table 2: Binary classification F1-score of different multimodal combinations of LSTM base models on a dialogue-level (worthwhile) label (Acoustic (A), Body (B), Visual (V), and Linguistic (L)).

	GRU (1-128)	LSTM(1-128)	LSTM(2-128)	LSTM(1-64)
L	0.677	0.691	0.66	0.618
L+A	0.601	0.67	0.664	0.56
L+B	0.597	0.605	0.664	0.563
L+V	0.683	**0.738**	**0.7**	0.694
ALL	**0.692**	0.728	0.618	**0.697**

5.1.1 Comparative experiment settings (single task on the dialog level): To investigate suitable hyperparameters, we first design comparative experiments to recognize dialog-level annotations.
GRU (1-128): A GRU layer with 128 hidden units is applied.
LSTM (1-128): An LSTM layer with 128 hidden units is applied.
LSTM (1-64)) An LSTM layer with 64 hidden units is applied.
LSTM (2-128): Two LSTM layers with 128 hidden units are applied.

For all the experiments, the number of epochs was 60, and we used the Adam optimizer with a learning rate of 0.001. Table 1 shows that the worthwhile label has the highest correlation coefficient (0.531) with self-sentiment. Therefore, we choose the worthwhile label as the target label. The numbers of high/low data points for the worthwhile labels at the dialog level were 38/22. The results of the comparisons are described in Section 6.1.

5.1.2 Multitask experiment settings. A 1-layer recurrent layer with 128 units was applied. The number of epochs was set to 60. In our experiments, we used the adaptive moment estimation (Adam) optimizer with a learning rate of 0.001. The information associated with various modes plays different roles in recognizing dialog-level labels and exchange-level sentiment. To determine the appropriate relationship between the two tasks in the multitask model, we used different values of λ {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}.

6 RESULTS

6.1 Comparison of different methods

Table 2 shows the results of 4 comparative experiments with the multimodal features. To obtain a stable model, we first use the GRU model and LSTM model with the same parameters (1 layer with 128 units) to recognize the worthwhile label. Columns 2 and 3 show the results of the GRU (1-128) and LSTM (1-128) models, respectively. The L+V feature set achieved the best result (0.738) with the LSTM(1-128) model, which is better than the best result (0.683) achieved by the GRU (1-128) model with the ALL feature set. For this reason, we use the LSTM model as the base model. Then, to obtain appropriate parameters, different parameter settings were applied in the LSTM model, and columns 3 to 5 present the LSTM

Table 3: Binary classification F1-score of different multimodal combinations of the LSTM (1-128) base model on a dialogue-level (worthwhile) label (Acoustic (A), Body (B),Visual (V), Linguistic (L)).

	Multitask									Single task
Loss weight (λ)	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	
L	0.648	0.603	0.668	0.702	0.615	0.677	0.722	0.668	0.629	0.691
L+A	0.697	**0.753**	**0.762**	0.639	0.677	0.707	0.715	0.663	**0.753**	0.67
L+B	0.639	0.715	0.692	0.625	0.657	0.563	0.59	0.643	0.601	0.605
L+V	0.615	0.7	0.715	0.643	0.677	**0.728**	**0.817**	0.713	0.707	**0.738**
ALL	**0.702**	0.744	0.677	**0.775**	**0.722**	0.629	0.775	**0.722**	0.692	0.728

(1-128), LSTM (2-128), and LSTM (1-64) model results. The best results of the LSTM (1-128), LSTM (2-128), and LSTM (1-64) models were 0.738, 0.7, and 0.697, respectively. For the LSTM (1-128) model, the best result (0.738) was obtained with the L+V feature set. Thus, the LSTM (1-128) model was used as the baseline. Moreover, we found that in most cases, the ALL and L+V feature sets performed better than the unimodal feature set (L), while the L+A and L+B feature sets performed worse than the unimodal feature set (L) in all cases.

6.2 Comparison of the single task and multitask models

Section 6.1 shows that the LSTM (1-128) model achieves the best F1-score with the L+V feature set. According to this result, we applied the same setting, namely, an LSTM layer with 128 hidden units, in the multitask model, and the structure is described in Section 4.2.2. Columns 1 to 10 in Table 3 show the binary classification results of the multitask model with different modality features and loss weight values (λ). Column 11 in Table 3 presents the binary classification results of a single task with different modality features. For the single-task model, the results with the ALL and L+V feature sets were better than those achieved with the unimodal feature set (L), and the L+V feature set produced the best result (0.738). For the multitask LSTM model, the L+V feature set achieved the best F1-score (0.817) by using MTL with loss weight loss (λ = 0.7). The best results of all feature sets (L, L+A, L+B, L+V, ALL) with the multitask model performed better than those with the single-task model, with results of 0.702, 0762, 0.715, 0.817, and 0.775, respectively, which represent improvements of 0.011, 0.092, 0.113, 0.079, and 0.047, respectively. The recognition performance resulted in a large improvement, demonstrating that our multitask model can learn the relation between exchange-level sentiment and a dialog level label (worthwhile) and is thus useful and effective for recognizing dialog-level labels. Meanwhile, for most experiments, the L+B feature set performed worse than the unimodal featur e(L) , and we suspect that the body feature does not work well to predict the worthwhile label. This also explains that why the L+V feature set obtains the best result in Table 2 and 3. On the other hand, in some cases (λ = 0.1), the multitask model performed worse than the single-task model, which indicates that a suitable weight loss is important.

6.3 Results of 18 types of annotations

Table 4 shows the best binary classification result of 18 types of annotations at the dialog level with the LSTM (1-128) model. Columns 3 and 4 present the best F1-score for a single task and a multitask with 5 types of comparative multimodal feature combinations, respectively. We observed that the multitask model performed better than the single-task model in all annotations. Compared with the average positive dialog annotation, the average negative dialog annotation was worse on the single-task model, with results of 0.701 and 0.685, respectively. After MTL was applied, the average negative dialog-level annotation (0.780) performed better than the average positive dialog-level annotation (0.767). Among all annotations, the cold label obtained the minimum F1-score (0.583) at the exchange level, and this label achieved the highest improvement (0.157) by applying MTL. The awkward label obtained the best F1-score (0.833) with the multitask model.

Table 4: Binary classification F1-scores of 18 annotations at the dialogue level. (a) shows positive dialog-level annotations; (b) shows negative dialog-level annotations. "diff" denotes the difference in F1-scores between the single task and multitask models.

(a)

	High/Low	Best MTL	Best Single	Diff
Well-coordinated	38/22	0.78	0.668	+0.112
Cooperative	46/14	0.737	0.683	+0.054
Harmonious	34/26	0.782	0.653	+0.129
Engrossing	27/33	0.798	0.71	+0.087
Involving	34/26	0.729	0.653	+0.076
Friendly	11/49	0.781	0.733	+0.048
Active	39/21	0.715	0.707	+0.008
Positive	46/14	0.798	0.76	+0.038
Worthwhile	38/22	0.817	0.738	+0.079
Average	/	0.767	0.701	+0.066

(b)

	High/Low	Best MTL	Best Single	Diff
Boring	16/44	0.778	0.767	+0.011
Unsatisfying	14/46	0.758	0.683	+0.075
Uncomfortably paced	35/25	0.826	0.679	+0.147
Cold	15/45	0.74	0.583	+0.157
Awkward	32/28	0.833	0.683	+0.150
Unfocused	18/42	0.795	0.644	+0.151
Intense	23/37	0.762	0.722	+0.040
Dull	15/45	0.744	0.705	+0.039
Slow	33/27	0.783	0.7	+0.083
Average	/	0.780	0.685	+0.095

Table 5: Binary classification F1-scores of the awkward and well-coordinated labels.

Label	Model	L	A+L	B+L	F+L	ALL	Human model(Wizard)
Well_coordinated	MTL	0.738	0.791	0.733	0.733	0.791	0.72
	Single	0.697	0.722	0.597	0.753	0.764	
	Multi-other [23]	0.7	0.74	0.65 (B+V+L)		0.76	
Awkward	MTL	0.733	0.744	0.744	0.766	0.783	0.58
	Single	0.649	0.649	0.648	0.75	0.633	
	Multi-other [23]	0.66	0.58	0.68 (B+V+L)		0.63	

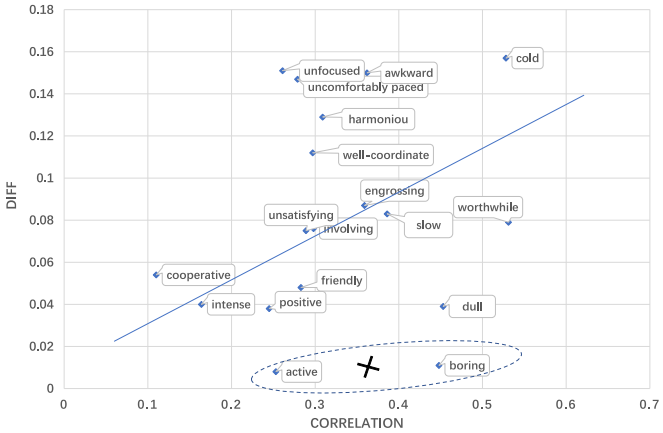

Figure 5: Analysis of the effect of correlations. (Diff denotes the difference in F1-scores between the single task and multi-task models, while Correlation shows the Pearson correlation coefficient between self-sentiment and dialog-level annotations.)

7 DISCUSSION

7.1 Comparisons with previous works

To the best of our knowledge, [23] used the same dataset as this work, which used LSTM (2-128) to recognize the awkward and well-coordinated labels. To compare our results with the multi-other model, we used the LSTM (2-128) model and applied MTL to recognize the awkward and well-coordinated labels. Table 5 shows the binary classification results of the awkward and well-coordinated labels. For the well-coordinated label, the single task achieves a similar result to that presented in [23]. All models achieved the best F1-score with the ALL feature set, and the multitask model produced the best result (0.783). For the awkward label, when comparing the single-task model with the multi-other model presented in [23], the results with most feature sets are similar, except for the A+L feature set. The multitask model achieved a better F1-score than all other feature sets (A, L+V) for the awkward label. Overall, the multitask model proposed in this work achieved a better performance on all feature sets than the results presented in [23], which demonstrates that our proposed method better utilizes exchange information and improves model performance. Column 8 shows the results of the human model proposed in [23]. The user satisfaction label score was annotated by Wizard, and the users were divided into high and low categories before the F1-score with the original annotation was calculated. Compared with the human model, the MTL models achieved better F1-scores on both the well-coordinated and awkward labels.

7.2 Analysis of the effect of correlations

By combining the correlation coefficients in Table 1 and the improvement results in Table 4, we found that the correlation coefficients and improved performance are not significantly related for some labels. For positive labels, the friendly label has a correlation coefficient of 0.253, while the multitask and single-task models produce almost the same result. For negative labels, the boring label has a coefficient -0.448 , only achieve a slight improvement(0.011) by using MTL. Meanwhile, we found that the absolute value of the average correlation coefficient of the negative labels (0.352) was higher than the average correlation coefficient of the positive labels (0.299), and the average improvement in the negative labels (0.095) by using MTL was higher than the average improvement in the positive labels (0.066), which indicates that although the performance improvement was not related to the correlation coefficient for every label, the correlation coefficient has a positive relationship with the overall improvement by comparing the positive labels and negative labels performances. Furthermore, we drew the scatter plot of the correlation coefficient between 18 types of labels and exchange-level sentiments, as well as the different improvements in the F1-score by using multiple tasks as shown in Fig.5. We observed that the correlation coefficient is positively correlated with the performance improvement for overall user impression.

8 CONCLUSION

In this paper, we first investigate the relationship between the exchange level and 18 dialog-level annotations. The worthwhile label has the highest correlation with user self-sentiment. To capture these correlations, we propose a multitask model to learn the relevant information. By comparing our proposed multitask model with a single-task model and other relevant research, we show that the multitask model achieved the best performance, with a 15.7% performance improvement over the signal-task model with cold labels. Thus, our results demonstrate that our model can utilize this relation to achieve better performance. However, there is still room for improvement. [19] indicates that sex information is beneficial for recognizing emotions. This study used only user modal information to recognize the users' impression of the dialog system, while other user characteristics, such as age and sex, were not considered. In future work, we will utilize the effects of user characteristics on user impressions to evaluate dialog systems.

ACKNOWLEDGMENTS

This work was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (Grant Numbers 22H04860 and 22H00536) and JST AIP Trilateral AI Research, Japan (Grant Number JPMJCR20G6). This work was also partially supported by the Research Program of "Five-star Alliance" in "NJRC Mater & Dev.".

REFERENCES

[1] Yaser S Abu-Mostafa. 1990. Learning from hints in neural networks. *Journal of complexity* 6, 2 (1990), 192–198.
[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. 59–66.
[3] Frank J Bernieri, John S Gillis, Janet M Davis, and Jon E Grahe. 1996. Dyad rapport and the accuracy of its judgment across situations: a lens model analysis. *Journal of Personality and Social Psychology* 71, 1 (1996), 110.
[4] Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019. Multi-domain conversation quality evaluation via user satisfaction estimation. *arXiv preprint arXiv:1911.08567* (2019).
[5] Praveen Kumar Bodigutla, Aditya Tiwari, Josep Valls Vargas, Lazaros Polymenakos, and Spyros Matsoukas. 2020. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. (2020).
[6] Joseph N Cappella and Sally Planalp. 1981. Talk and silence sequences in informal conversations III: Interspeaker influence. *Human Communication Research* 7, 2 (1981), 117–132.
[7] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. 2007. Observer-based measurement of facial expression with the Facial Action Coding System. *The handbook of emotion elicitation and assessment* 1, 3 (2007), 203–221.
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. (2018), 4171––4186.
[9] Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling user satisfaction with hidden Markov models. In *Proceedings of the SIGDIAL 2009 Conference*. 170–177.
[10] Florian Eyben and et al Scherer, Klaus R. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2015), 190–202.
[11] Sunao Hara, Norihide Kitaoka, and Kazuya Takeda. 2010. Estimation method of user satisfaction using n-gram-based dialog history model for spoken dialog system. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.
[12] Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Modeling user satisfaction transitions in dialogues from overall ratings. In *Proceedings of the SIGDIAL 2010 Conference*. 18–27.
[13] Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. 2019. Multitask prediction of exchange-level annotations for multimodal dialogue systems. In *2019 International Conference on Multimodal Interaction*. 85–94.
[14] Michimasa Inaba, Naoyuki Iwata, Fujio Toriumi, Takatsugu Hirayama, Yu Enokibori, Kenichi Takahashi, and Kenji Mase. 2014. Constructing a Non-task-oriented Dialogue Agent using Statistical Response Method and Gamification.. In *ICAART (1)*. 14–21.
[15] Yohei Kikuta. 2019. BERT Pretrained model Trained On Japanese Wikipedia Articles. https://github.com/yoheikikuta/bert-japanese.
[16] Kazunori Komatani and Shogo Okada. 2021. Multimodal Human-Agent Dialogue Corpus with Annotations at Utterance and Dialogue Levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.
[17] Gregory J McHugo, John T Lanzetta, Denis G Sullivan, Roger D Masters, and Basil G Englis. 1985. Emotional reactions to a political leader's expressive displays. *Journal of Personality and Social Psychology* 49, 6 (1985), 1513.
[18] Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America* 119, 4 (2006), 2382–2393.
[19] Dinh Viet Sang, Le Tran Bao Cuong, and Vu Van Thieu. 2017. Multi-task learning for smile detection, emotion recognition and gender classification. In *Proceedings of the Eighth International Symposium on Information and Communication Technology*. 340–347.
[20] Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. 149–152.
[21] Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. 2011. Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*. 173–184.
[22] Stefan Ultes. 2020. Improving interaction quality estimation with BiLSTMs and the impact on dialogue policy learning. *arXiv preprint arXiv:2001.07615* (2020).
[23] Wenqing Wei, Sixia Li, Shogo Okada, and Kazunori Komatani. 2021. Multimodal User Satisfaction Recognition for Non-task Oriented Dialogue Systems. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 586–594.
[24] Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proc. IEEE* 101, 5 (2013), 1160–1179. https://doi.org/10.1109/JPROC.2012.2225812
[25] Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alexander Rudnicky. 2016. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *Proceedings of the 17th annual meeting of the Special Interest Group on Discourse and Dialogue*. 55–63.
[26] Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*. 404–412.