Task-independent Multimodal Prediction of Group Performance Based on Product Dimensions

Go Miura
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
s1810175@jaist.ac.jp

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

ABSTRACT

This paper proposes an approach to develop models for predicting the performance for multiple group meeting tasks, where the model has no clear correct answer. This paper adopts "product dimensions" [Hackman et al. 1967] (PD) which is proposed as a set of dimensions for describing the general properties of written passages that are generated by a group, as a metric measuring group output. This study enhanced the group discussion corpus called the MATRICS corpus including multiple discussion sessions by annotating the performance metric of PD. We extract group-level linguistic features including vocabulary level features using a word embedding technique, topic segmentation techniques, and functional features with dialog act and parts of speech on the word level. We also extracted nonverbal features from the speech turn, prosody, and head movement. With a corpus including multiple discussion data and an annotation of the group performance, we conduct two types of experiments thorough regression modeling to predict the PD. The first experiment is to evaluate the task-dependent prediction accuracy, in the situation that the samples obtained from the same discussion task are included in both the training and testing. The second experiments is to evaluate the task-independent prediction accuracy, in the situation that the type of discussion task is different between the training samples and testing samples. In this situation, regression models are developed to infer the performance in an unknown discussion task. The experimental results show that a support vector regression model archived a 0.76 correlation in the discussion-task-dependent setting and 0.55 in the task-independent setting.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**.

KEYWORDS

Group performance, Multimodal, Group Analysis

ACM Reference Format:
Go Miura and Shogo Okada. 2019. Task-independent Multimodal Prediction of Group Performance Based on Product Dimensions. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340555.3353729

1 INTRODUCTION

The computational analysis of group conversations has enormous value for social sciences, and could be important for implementing relevant applications that support interaction and communication, including self-assessment, training and educational tools, and systems to support group collaboration through the automatic sensing, analysis, and interpretation of social behavior [11]. As one application, identifying factors of a meeting process and predicting the output of a group are the central challenges of group meeting analysis. Although there is great interest in the notion of performance prediction in science and real-life applications, one of the main difficulties in such tasks is due to the wide variety of definitions of performance and discussion types. In recent research [24], [2], the performance of the group output is defined as the difference between the answer after group decision making and the correct answer by experts. In addition, these studies focus on group performance modeling on one discussion task. However, the exact correct answer in a group meeting is not always defined, and the group needs to solve various types of discussion tasks in a real life. Toward developing prediction models of outputting the performance for multiple group meeting tasks that do not have a clear correct answer such as in business meetings, another

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3353729

type of evaluation metric measuring the group output is also required.

As an annotation method, we adopt "product dimensions" (PD) [13], which is proposed as a set of dimensions for describing the general properties of written passages that are generated by the group. The product dimensions are composed of a total 18 degrees based on six dimensions: Action Orientation, Length, Originality, Optimism, Quality of Presentation, and Issue Involvement. These 18 independent degrees cover various types of views of performance such as the issue of involvement, originality of ideas, and being appropriate as a reference for a third party. In this study, we annotate the PD to measure the group performance and present a computational analysis of the prediction of the PD.

We regard written passages (outcome of discussions) as a transcript (discussion log) of the group discussions. The transcript includes all utterances spoken by the group members, and the annotators can judge the contents of the discussions and how many ideas or what types of ideas are outputted in the discussions by observing the transcript. The performance is annotated based on the relationship between (1) the objective and subject of the group meeting task and (2) the manuscript of the discussion as a group output.

For this study, we use the group discussion corpus proposed in [25], called the MATRICS corpus. The corpus includes 30 group discussion sessions by 40 participants, for which the total discussion length is more than 9 hours. We enhance the MATRICS corpus by annotating the performance metric of the PD. The product dimensions were assessed by 4 external coders. The corpus includes three types of discussion sessions performed by each group.

From each corpus, we extract group level linguistic features including vocabulary level features using the word 2 vec technique, topic segmentation techniques, and functional features with dialog act and parts of speech on the word level. We also nonverbal features from the speech turn, prosody, and head movement. We develop regression models using the multimodal features for each discussion task and analyze which features are effective for the prediction of the group performance. The task-independent performance prediction, which is a novel prediction task, is then conducted. We summarize the main contributions in this work.

Group performance defined by product dimensions: As a group performance index, the usage of product dimensions is unexplored on computational analysis. We set a novel challenge for group performance prediction with this new index. Product dimensions capture various types of views of the team performance. The multimodal modeling of the dimensions enable us to analyze the differences in the effective multimodal features to predict the performance in each discussion task. The details of the annotation are described in Section 3.

Task-independent performance prediction: Many previous studies conducted the group output prediction task with a single discussion type. The MATRICS corpus includes three types of discussion sessions done by each group. This study presents a computational analysis to predict the task-independent group performance with the multiple group discussion corpus. The results of the task-independent group performance are described in Section 6.

Various types of linguistic features: This research shows various types of linguistic features including the vocabulary, topic transitions, parts of speech, and dialog act enable us to analyze the differences in the group performance. The contributions of linguistic features are described in Section 7 and Tables 5 through Section 6.

2 RELATED WORKS

group meetings have been studied in computer science. The goal is to develop a method that automatically analyzes the interaction process using both the spoken words and the nonverbal behaviors [12], [11]. According to [12], the challenge is classified into four groups: (1) analysis of conversational dynamics, (2) verticality and roles, (3) personality traits, and (4) group-level analysis. The computational analysis of group performance is belong to (4) group level analysis. Woolley et. al, [35] presents a quantitative analysis of the collective intelligence in a group and the intelligence of the group members using a dataset collected from 107 groups. The findings show that the collective intelligence in the group performance is correlated with the communication patterns, including turn taking. Dong et. al, [9] presented an analysis of the group performance in a brainstorming task. Dong et. al, [10] analyzes the between group performance and the communication patterns based on turn taking by using a mixture of hidden Markov processes (HMPs) in learning the structure of group dynamics from nonverbal audio features. Hung et. al, [14] presents an approach for estimating the cohesion in group by using several turn-taking features and audio-visual features.

For the group meeting analysis, recent research projects released and shared group meeting corpora for analyzing human-human multimodal conversations, such as the Augmented Multiparty Interaction (AMI) corpus [5] and the ICSI meeting corpus [15]. The Computers in Human Interaction Loop (CHIL) corpus considers human-human interactions in offices and classrooms [34], and the Video Analysis and Content Extraction (VACE) corpus treats human-human interactions in battle-game sessions in the air force [7]. In recent years, the Emergent LEAder corpus (ELEA) which is a multiparty multimodal dataset that allows gathering information on the group performance, dynamics, and structure [30], was released. In this corpus, the winter survival task is a fictional scenario, the purpose of which is to rank 12 items in order to

survive an airplane crash in the winter. For each group, the ranking was done individually by participants at the beginning of the task and as a team after the group discussion to elucidate the effects of cooperation and factors determining the group composition, e.g., dominance and leadership. The individual and group rankings were then compared with the rankings of survival experts, and Absolute Individual Scores (AIS) and Absolute Group Scores (AGS) were calculated via absolute differences in rank order, as measures of the individual and group performance, respectively. The ELEA corpus contains individual-level annotations of personality traits and leadership and group-level annotations including group performance, so the corpus is used for leadership modeling [29], and personality trait modeling [1], [21].

The ELEA corpus is also used for group performance analysis and modeling [16], [2], [24]. Jayagopi et al. [16] proposed a method to extract group level nonverbal patterns by using latent Dirichlet allocation (LDA)[4] and then investigated these group nonverbal patterns and group performance or group composition. Avci et al. [2] models the influence of one member on the other members by relating the interactions of nonverbal patterns between group members to the transition between hidden states (e.g., one utterance starts after an utterance by another member) in a Markovian formulation. The findings in [2] show that the interaction features of nonverbal patterns are effective to improve the prediction accuracy of the group performance in ELEA. Murray et al. [24] presented a prediction model using speech features and linguistic content features that are not used in [2]. In addition, two approaches: transfer learning and the data augmentation method, are adopted for increasing the amount of small size of group meeting data. The experimental results show that both approaches improve the prediction accuracy of the group performance in ELEA. From the findings using ELEA corpus, though a winter survival task in the ELEA corpus is one of the effective types of discussion to analyze the group performance because the correct-answer can be defined by an expert, in general, the discussion tasks vary widely, and the CORRECT-ANSWER by an expert does not exist in many discussion cases.

Studies [6] [31] [19] on social science investigated how to define the group performance based on various criteria, such as the kind of group activities in the task, the relations among the group members and the kind of performance processes in the task, and so on. Hackman took a different approach to classifying the task differences and relating them to the group performance [13]. For this research, Hackman proposed an index of the group performance called "product dimensions". In the evaluation process using the product dimensions, the output of a task is defined as a product, and the products are assessed by the dimensions. We used Hackman's six product dimensions because these dimensions can be applied for written product and we consider that these dimensions are appropriate for the task-independent performance prediction.

We annotated the performance based on the product dimensions into 29 discussions in the MATRICS corpus. The MATRICS corpus with the product dimensions annotation enable us to address a novel challenge on the computational analysis of task-independent group performance. We develop regression models to predict the product dimensions using group-level multimodal nonverbal features and evaluate the task-independent prediction accuracy, i.e., the accuracy by which a model correctly predicts the performance of a discussion task that is not used as the training data. In addition, we show the novel linguistic features capturing topic transitions in a discussion, which are extracted using a topic segmentation technique [23]. The linguistic features of the dialog act in group discussion have been unexplored in previous works such as [24].

As related works using the MATRICS corpus, Nihei et al. [25] investigate the influential statements that affect the discussion flow and are highly related to facilitation, and develop a model that predicts influential statements in group discussions. Nihei et al. [26] proposed deep learning framework to improve the prediction accuracy. Okada et al. [27] presents a computational analysis of individual communication skills and investigates the relationship between the discussion task type and the effective multimodal features in estimating the skills. The group performance prediction, though was not the focus in [25], [27], [26].

3 GROUP DISCUSSION CORPUS

We provide an overview of the MATRICS corpus (MultimodAl (Task-oriented) gRoup dIsCuSsion) corpus) [25] used in this study. The corpus is collected to analyze the discussion process in the group discussion (GD) and analyze the communication skills of the group members [27]. 40 participants are recruited to develop the corpus through a human resource company. Each group is composed of four participants, so that a total of 10 groups are composed. In each group, none of the participants had previously met each other. Each group had three discussion sessions in Japanese, with one task being addressed in each session. To cancel out the effect of the task order, the order of the tasks was randomized. The three types of tasks are as below:

Task 1: Celebrity guest selection (in-basket): The participants were asked to pretend that they were the executive committee members for a school festival, and were choosing a guest to invite to the festival. The participants engaged in a discussion to determine the ranked order as a group.

Task 2: Booth planning for a school festival (case study): The participants were instructed to discuss and create a plan for a small booth to sell food or drinks at a school festival.

Task 3: Travel planning for foreign friends (case study): The participants were instructed to create a two-day travel plan for foreign friends who are visiting Japan on vacation.

In this study, we used audio data obtained from a head set microphone, head acceleration data obtained from acceleration sensors and manual transcription data of spoken utterances in the MATRICS corpus. These data are used to extract multimodal features including linguistic, motion, and acoustic features.

Definition and Analysis of Group Performance

We use the index of Product Dimensions [13] defined by Hackman as a task-independent index of the group performance. Hackman analyzed the relationship between the task type and group performance by using the product dimensions. The advantage that the index is available to measure the performance for different types of tasks is useful for the purpose of this research.

Table 1 shows this index. The indices can be applied to written passages. We defined the manual transcript of spoken utterances as the output of the GD as the written passage in this research, because third-party coders can judge how many ideas are appear, and how the group conclude the discussion and output the answer by observing the transcripts. According to [13], the 18 indices are composed of six groups as follows:

(I) Action orientation: The degree to which a product states or implies that a specific or general course of action should be, might be, or will be followed.

(II) Length: The degree captures the length of the written passage including the number of words, number of adjectives, shortness (negative), and lack of detail and elaboration (negative).

(III) Originality: The degree to which the ideas and/or mode of presentation of a product are fresh and unusual as opposed to obvious and mundane.

(IV) Optimism: The degree to which the general point of view or tone of a product can be characterized as "positive" or optimistic as opposed to "negative"or pessimistic.

(V) Quality of presentation: Evaluation of the grammatical, rhetorical, and literary qualities of the product.

(VI) Issue involvement: The degree to which a product takes or implies a particular point of view regarding some goal, event, value or procedure.

Four external coders observe a pair of (1) transcripts of discussion and (2) the instruction of the discussion task. These coders annotated the Product Dimensions. We defined the total score annotated by the 4 external coders as the group performance (training label of machine learning). The index of the scale is from 1 to 7. The number of discussions in the sample is 30, but we use 29 samples for the experiments because the data of one group could not be recorded due to machine trouble. (R) denotes the reverse index; if the value is large, it means more negative.

Table 1: Krippendorff's alpha (κ_k) for each dimension (The column "Selected" indicate we use or not use the scales in this study. From "(I)-C" to "(VI)-P" indicate used scales in this study, "-" indicates not use.)

Dimension	Descriptive scales	Selected	κ_k
(I) Action Orientation	(I)-C: Constructive	(I)-C	0.30
	Suggests action	-	0.08
	Passive (R)	-	0.15
(II) Length	Lacks detail, elaboration (R)	-	0.30
	Short (R)	-	0.16
(III) Originality	(III)-O: Original	(III)-O	0.27
	(III)-B: Bizarre	(III)-B	0.28
	(III)-N: Not unusual (R)	(III)-N	0.27
(IV) Optimism	Disapproves (R)	-	-0.01
	(IV)-P: Positive outlook	(IV)-P	0.43
	Supportive	-	0.12
(V) Quality of Presentation	(V)-U: Understandably presented	(V)-U	0.29
	Choppy (R)	-	0.19
	(V)-S: Stylistically well-integrated	(V)-S	0.20
(VI) Issue Involvement	Low issue involvement (R)	-	0.16
	(VI)-S: States a belief or opinion	(VI)-S	0.32
	(VI)-P: Propagandistic	(VI)-P	0.20

Table 2: Correlation between dimensions (The descriptions of "(I)-C" to "(VI)-P" are shown in Table 1)

Scales	(III)-O	(III)-B	(III)-N	(IV)-P	(V)-U	(V)-S	(VI)-S	(VI)-P
(I)-C	0.35	0.16	-0.18	0.72	0.74	0.63	0.40	0.55
(III)-O		0.80	-0.71	0.41	0.20	0.03	0.74	0.72
(III)-B			-0.93	0.27	-0.04	-0.19	0.78	0.62
(III)-N				-0.24	0.06	0.20	-0.74	-0.54
(IV)-P					0.73	0.48	0.34	0.46
(V)-U						0.72	0.03	0.23
(V)-S							0.00	0.22
(VI)-S								0.87

We calculated the level of agreement (Krippendorff's alpha[18]) for each index between four coders. Table 1 shows the Krippendorff's alpha for each index. These values is lower than 0.5, so we did not get sufficient agreement. It seems that the annotation of the group performance is subjective, so a high score agreement was not obtained. In this study, we ignore 7 indices. Suggests action, Passive (R), Short (R), Disappoves (R), Supportive, Choppy (R), and Low issue involvement (R), for which the Krippendorff's alpha is less than 0.2.

In addition, we also except all indices in (II) Length from the research subjects, because the degree of (II) Length of the discussion correspond to the length of the meeting or size of vocabulary, and these indices can be obtained by a simple and robust method such as the counting the utterances rather than requiring a multimodal group discussion analysis. The remaining 9 indices are the target variables for this study. The 9 indices are shown the column "Selected " in table 1. We performed a correlation analysis between the annotated scores by the external coders and the product dimensions

to investigate the agreement between coders. Table 2 shows the Spearman's rank correlation coefficient among the 9 indices in product dimensions. Table 2 shows that correlation coefficient of the descriptive scales in the same dimension index is more than 0.7 (for the reverse index, less than -0.7). The table shows that the scores of the indices in a dimension group have a high similarity to each other. For simplicity, we calculated the mean value of the score based on these scales in the same dimension index, and unite some scales into one dimension index. Consequently, the 9 descriptive scales are united into 5 dimension indices. Additionally, we use the reverse score for (III) Not unusual (R).

4 EXTRACT MULTIMODAL FEATURES

From the findings from [24],[2], the multimodal (visual and acoustic) nonverbal features and linguistic features are effective to predict the performance. In addition, the scores of the Product Dimensions are annotated by referring to the transcript, and so that it is important to extract the group-level linguistic features as the key descriptors to predict it. For this study, the individual features that are used to calculate the group-level features are partly shared with [27].

Speaking Turn Features

Total speaking length: The total speaking length is calculated as a summation of the length of the utterances observed in a session.
Total count of utterances: The count is calculated by counting the utterances observed in a session.
Total speaking length ($\geq$ 1s): The length is calculated as a summation of the length of the utterances that is longer than 1 second.
Total count of utterances ($\geq$ 1s): The count is calculated by counting the utterances that are more than 1 second.

Acoustic Features

The acoustic features are extracted from the speech signals of an utterance segment. An audio processing tool Praat[1] and openSMILE [2] software are used to extract the acoustic features. The intensity level is calculated to capture the loudness of the voice, and this is correlated with the energy value.
Maximum, minimum and mean pitch: The maximum, minimum and mean values of the pitch (MaxPitch, MinPitch, MeanPitch) are calculated in the utterance segments. The maximum, minimum and mean values of the pitch in a session are defined as the mean values of the MaxPitch, MinPitch, MeanPitch calculated from all the utterances observed in a session.
Maximum and minimum intensity: These features (MaxIntensity, MinIntensity) are calculated in the same manner as MaxPitch, and MinPitch.
Difference in intensity: The feature is calculated as MaxIntensity - MinIntensity.
Difference in pitch: The feature is calculated as MaxPitch - MinPitch.
Speaking speed: The speaking speed sp_t is the number of syllables in an utterance t divided by the utterance length. The speaking speed sp is calculated by averaging sp_t over all the utterances.
Group-level acoustic features: We also extracted acoustic features including mfcc (Mel-frequency cepstral coefficients), pcm RMSenergy (Root mean-square signal frame energy), pcm zcr (Zero-crossing rate of time signal), voiceProb (The voicing probability computed from the ACF), and F0 (The fundamental frequency computed from the Cepstrum) per utterance observed in the discussion, using openSMILE software. The features per utterance are averaged over all utterances, and the averaged features are used as group-level features.

Motion: Head Activity

The features are calculated from three coordinate signals (x,y,z), observed from the sensor.
Mean and deviation of head movement: The norm $|a_t|$ of the acceleration : $a_t = \{x_t, y_t, z_t\}$ is calculated at time t. The mean and the standard deviation values of $|a_t|$ are calculated in a session to extract the feature corresponding to the amount of movement.
Mean and deviation of movement while speaking: Using the utterance segment that is calculated in extraction speaking turn features, the features of head activity are extracted while the participant was speaking. Let M utterance segments: $UT = \{ut_1 \dots ut_m \dots ut_M\}$ be observed in the session. After calculating the mean $|\bar{a_m}|$ and standard deviation $std|a|_m$ in ut_m, the mean $|\bar{a_m}|$, $std|a|_m$ in all utterances are calculated. The statistics were defined as the mean and the deviation of the movement while speaking.

Linguistic Features

Part of Speech (PoS): The number of words spoken per type of grammatical construction are counted. The word features were extracted from the transcripts using a Japanese morphological analysis tool: MeCab[17]. First, the sentence was segmented into word sets by the tool. Second, the type of the part of speech (PoS) is automatically annotated to each word in the word sets.
Noun, Verb, Injection, Filler: The spoken words for PoS: "Noun", "Verb", "Injection", "Filler" are counted.

[1] http://www.fon.hum.uva.nl/praat/
[2] https://www.audeering.com/opensmile/

New noun: First, the instant is identified at which a noun is spoken for the first time in the discussion and defined as a new noun. Second, the number of new nouns is counted per participant.

Dialog Act and Semantic Tag: A speech act is a primitive abstraction of the typically illocutionary force of utterance [33]. However, it ignores the conversational aspect of the spoken interaction and contributions. To analyze the discourse structure and understand the conversational function, a dialog act tagging scheme is proposed in the literature [8][32]. In this study, the annotation data of dialog act tags and semantic tags were collected in [27].

Definition of tags: These tags are annotated into each utterance by referring to the manual transcription. The ten types of tag are "Conversational opening", "Open question", "Suggestion", "Backchannel", "Open opinion", "Partial accept", "Accept", "Reject", "Other questions" from the DAMSL (Dialog Act Markup in Several Layers) tag set [8] and "Understanding Check" from the MRDA (Meeting Recorder Dialog Act) tag set [32]. "Other questions" is a tag merging all question tags such as "WhQuestion" and "Y/N question" except for "Open question". In addition, "Plan", "Agreement" and "Disagreement" are annotated as a speech act tags. Two semantic tags, "Describe fact" and "Reason" are annotated. A describing fact is often observed in the statements of the participants. When the participant gives his/her opinion, giving the reason at the same time is useful to emphasize the opinion. Given these reasons, we set the semantic tags. We set a tag of "Interjection" because utterances including only interjection such as "Well" and "So, you see" are also often observed and the other utterances are categorized into "Other utterance". Additionally, "Affirmation" tag is annotated, so a total of 19 tags are used.

Bigram Features of Tags (Dialog act): The bigram features are calculated based on the adjacency utterances. 361 (19×19) bigram patterns are created. After removing those patterns with a low frequency (less than 10) from the summation of the pattern set of each session, 98 bigram patterns remain as a feature set.

Vocabulary Level Features: In this research, we extract linguistic features based on a word embedding method for analyzing the contents of the vocabulary in the GD. An objective of word embedding is to learn feature representation, where words are mapped to vectors of real numbers. It is also used for dimension reduction from a space with many dimensions per word to a continuous vector space with lower dimension. Each word in an utterance is input to the word embedding model that is pre-trained using a word2vec technique [22] [20]. The word-embedding model was trained using the Japanese Wikipedia corpus. The feature extraction procedure is as follows: Vector: V_{w_i} in the (embedded) vector space of word w_i by using the model is averaged over all words in a utterance as the utterance vector V_u.

Vocabulary: We calculated the average of each element in V_u for all spoken utterances. The dimension of the vector space is set to 200, so the number of features is 200, which corresponds to the dimension of the vector space.

Topic transition (Topic): The topic segmentation is effective to extract content features for opinion mining or sentiment analysis [23], because how the topic changes is correlated to the opinion type or sentiment type. From this back ground, we utilize the topic segmentation technique to capture the topic transition in a discussion, and the features are used for predicting the group performance. The feature extraction process by a basic topic segmentation method is as follows.

First, we sort the utterances of all the participants in a group into time-series order. Let the mean utterance vector be averaged from the t th utterance vector: V_{u_t} to $t+F$ th vector: $V_{u_{t+F}}$ is MV_{u_t}. Second, we calculate the cosine similarity: $S(t, t+1)$ between consecutive mean-utterance vectors: MV_{u_t} and $MV_{u_{t+1}}$. We calculate the $S(t, t+1)$ in $t = 1, \ldots, T-F$ and extract features based on the time-series similarity data S, where F is smoothing parameter for moving average The time-series S is segmented using threshold Th and the peak detection is conducted where similarity $S(*, *)$ is less than threshold Th. Number of peaks captures change of distribution of vocabulary. We calculate maximum, minimum, maximum - minimum, mean, and standard deviation of S as topic transition features. We extracted these features with changing the segmentation threshold Th in $[0.6, 0.7, 0.8]$ and the parameter F in $[6, 7, 8]$.

Group-based feature extraction

We transform the individual features of speaking turn, acoustic, linguistic, and motion features to group features. 4 feature sets, Group-level acoustic features, Topic, Dialog act, and Vocabulary, have bee n extracted as group-level features by averaging utterance-level features. Except for these 4 feature sets, we calculated the maximum, minimum, differences of the maximum and minimum, mean, standard deviation, and summation of the individual features of 4 participants in order to transform them to group features. The number of individual features before the transform is 21. We calculated the group values of these features, and that of group features is 126. To summarize the dimensions, the number of Group-level acoustic features is 816, that of transformed speaking turn features is 24, that of transformed acoustic features is 24, that of transformed PoS is 30, that of Dialog act is 98, that of Vocabulary is 200, that of Topic is 21, and the total dimensions of the features is 1213.

5 EXPERIMENTS

Experiments are conducted to evaluate the prediction accuracy of the group performance (product dimensions).

$ **Regression prediction:** We used the Spearman's correlation coefficients r as a criterion for the performance of the regression prediction. For the regression task, we use the support vector regression (SVR) model using the RBF kernel. The ϵ-insensitive error is used as a loss function in SVR. These parameters of SVR are optimized using a nested cross-validation scheme in a training data set, with ϵ that adjusts the tolerance error selected from [0, 0.01, 0.1, 1] , γ from $[1/TN, 1/(TN \times Var_T)$, 0.01, 0.1, 1, 5, 10], where TN denotes the number of training features and Var_T denotes the standard deviation value of training features, and the penalty parameter C from [0.01, 0.1, 1, 5, 10].

$ **Feature selection:** We calculate the Spearman's correlation coefficient r and p value between training features and the training label. We select only features with a p value less than 0.1 and use them for training.

$ **Task-dependent or -independent prediction tasks:** We conducts two types of experiments. In the first experiment, we evaluate the regression accuracy of the performance score by leave-one-group-out testing. In this experiment, three samples corresponding to three discussion tasks in one group are used for testing. This setting is regarded as a task-dependent prediction task because the samples from same type of task are included in both the training and test data. The task setting in [24],[2] is also task-dependent, because the ELEA corpus with one discussion task is used for the evaluation. In second experiment, we evaluate the regression accuracy of the performance score by leave-one-sample-out testing, where the task of the testing sample is not used for training. This setting is regarded as a task-independent prediction task, and it is an unexplored task in previous works.

6 RESULTS

We compare the accuracy with the 15 feature set to analyze the contribution of each feature set for the prediction accuracy. "Acoustic", "Speaking turn", "Linguistic", "Motion" are represented as A, S, L, and M, respectively. Additionally, the multimodal features are combinations of the four modal features. For example, "Acoustic and Speaking turn" is represented as $A + S$.

$ **Task-dependent Prediction Results:** Table 3 shows the regression prediction accuracy of models. The values in table 3 denote the Spearman's correlation coefficient r. The bold values indicate the best accuracy for the performance index. The underlined values indicate that the correlation has a significant positive correlation where $p < 0.01$ and $r > 0.468$. For (III) Originality, the "acoustic + speaking turn + linguistic" ($A + S + L$) model and the model with all features (All) obtained the best accuracy: 0.76. In addition, the unimodal models A and L and all multimodal models except $S + M$ obtained significant correlations ($r > 0.468$). The overall prediction accuracy for (III) Originality is the highest in all performance indices. For (IV) Optimism, the "language + motion" ($L + M$) model obtained the best accuracy: 0.51 with significant correlation. It is only one case that the best model includes motion features. It is found that the head activity is also effective to capture the (IV) Optimism. For (VI) Issue Involvement, the best models are same as those for (III) Originality, and the accuracy is 0.54. For (I) Action Orientation and (V) Quality of Presentation, no significant correlation could be obtained. For all indices, The multimodal fusing is effective to improve the accuracy of the best unimodal model. The maximum improvement is 0.17 for (VI) Issue Involvement.

$ **Task-independent Prediction Results:** Table 4 shows the regression prediction accuracy (Spearman's correlation) of the task independent models. The values in table 4 denotes the mean value of the Spearman's correlation coefficients r in three tasks. Column 2 "Best" denotes the best accuracy in the task-dependent experiments (Table 3). The best accuracies in the task-independent experiments were worse than those in the task-dependent experiments for almost performance indices except (VI) Issue Involvement. It is natural that the discussion tasks where training samples are observed are different from the task where the testing sample is in the task-independent experiments, so that the distribution of the training samples is different from that of the testing samples.

For (III) Originality and (VI) Issue Involvement, the "speaking turn" (S) model obtained the best accuracies, 0.51, and 0.55, respectively, with significant positive correlations. Best accuracy 0.55 for (VI) Issue Involvement is better than that (0.54) for (VI) on the task-dependent task. The result means the performance of (VI) Issue Involvement could be predicted without depending on the type of task. For (IV) Optimism, the "language" (L) model obtained the best accuracy: 0.42. From these results, multimodal fusing is not effective on task-independent prediction task. This means that the effective multimodal combination is different depending on the task type. For (I) Action Orientation, the multimodal fusing of "acoustic + speaking turn" ($A + S$) slightly improve the best accuracy of unimodal features (0.36) by 0.01. The finding opens a new challenge of extracting common effective multimodal features and a multimodal fusion algorithm for the task-independent prediction setting. Conversely, the accuracy of the speaking turn or proposed linguistic features are robust to the different task types.

7 DISCUSSION

In both task-dependent and task-independent experiments, it is found that the linguistic features are effective to improving the performance prediction accuracy. In this section, we analyze the contributions of the linguistic features to predict the performance. The regression model is trained using linguistic features by removing the specific features: PoS, Topic,

Table 3: Regression prediction accuracy on task-dependent experiments (The accuracy denotes the Spearman's correlation coefficient r. The bold values indicate the best accuracy for the performance index. The underlined values indicate that the correlation has a significant positive correlation where $p < 0.01$)

Dimension index	A	S	L	M	A S	A L	A M	S L	S M	LM	A S L	A S M	A L M	S L M	All
(I) Action Orientation	0.31	-0.16	0.29	-0.11	**0.44**	0.35	0.40	0.31	0.09	0.38	**0.44**	0.36	0.40	0.39	0.43
(III) Originality	0.49	0.38	0.65	-0.74	0.65	0.72	0.49	0.62	0.38	0.65	**0.76**	0.65	0.72	0.62	**0.76**
(IV) Optimism	-0.06	-0.78	0.44	0.25	-0.18	0.25	-0.04	0.41	0.08	**0.51**	0.24	-0.11	0.29	0.40	0.30
(V) Quality of Presentation	0.19	-0.56	0.01	-0.25	0.12	0.12	**0.24**	-0.03	-0.39	-0.03	0.10	0.16	0.13	-0.09	0.10
(VI) Issue Involvement	0.25	0.37	0.37	-0.73	0.37	0.52	0.25	0.39	0.37	0.37	**0.54**	0.37	0.52	0.39	**0.54**

Table 4: Regression prediction accuracy on task-independent experiments (Spearman's correlation coefficient r, the bold values and the underlined values indicate the same functions as in Table 3, Column 2 "Best" denotes the best accuracy in the task dependent experiments (Table 3).)

Dimension index	Best	A	S	L	M	A S	A L	A M	S L	S M	L M	A S L	A S M	A L M	S L M	All
(I) Action Orientation	**0.44**	0.10	0.36	0.23	-0.01	**0.37**	0.20	0.13	0.26	0.31	0.24	0.24	**0.37**	0.18	0.26	0.22
(III) Originality	**0.76**	0.21	**0.51**	-0.04	-0.45	0.41	0.41	0.21	0.10	**0.51**	-0.04	0.42	0.41	0.41	0.10	0.42
(IV) Optimism	**0.51**	0.14	0.06	**0.42**	0.24	0.17	0.33	0.11	0.40	0.15	0.41	0.33	0.13	0.33	0.36	0.32
(V) Quality of Presentation	**0.24**	0.14	-0.10	0.12	-0.01	0.03	0.16	0.14	0.12	-0.15	0.10	**0.16**	0.02	0.14	0.07	0.14
(VI) Issue Involvement	**0.54**	0.22	**0.55**	-0.09	-0.28	0.36	0.17	0.22	-0.01	**0.55**	-0.09	0.23	0.36	0.17	-0.01	0.23

Table 5: Regression prediction accuracy (Spearman's correlation coefficients r) using linguistic features

Index	All_L	$All_L - D$	$All_L - P$	$All_L - T$	$All_L - V$
(I)	0.29	0.27 (+0.02)	0.21 (+0.08)	0.33 (-0.04)	0.04 (+0.25)
(III)	0.65	0.65 (±0.00)	0.67 (-0.02)	0.68 (-0.03)	0.37 (+0.28)
(IV)	0.44	0.43 (+0.01)	0.45 (-0.01)	0.41 (+0.03)	-0.09 (+0.53)
(V)	0.01	0.18 (-0.17)	0.08 (-0.07)	0.12 (-0.11)	-0.51 (+0.52)
(VI)	0.37	0.51 (-0.14)	0.35 (+0.02)	0.36 (+0.01)	0.14 (+0.23)
mean	0.35	0.41 (-0.06)	0.35 (±0.00)	0.38 (-0.03)	-0.01 (+0.36)

Dialog act, Vocabulary. Table 5 shows the regression accuracies of the model using linguistic feature sets that exclude specific features (e.g., $All_L - *$) for all indices, where P: PoS, T: Topic, D: Dialog act, and V: Vocabulary are substituted for $*$. Value x in (x) in the table denotes the difference in accuracy for the cases in which the feature set is removed.

From the table, V: Vocabulary feature set is most effective to improve the accuracy, because the mean difference of accuracy is +0.36. "Dialog act" contributes to improving the accuracy for (I) Action Orientation and (IV) Optimism by +0.02, +0.01. "PoS" contributes to improve the accuracy for (I) Action Orientation and (VI) Issue Involvement by +0.08, +0.02. "Topic", which is the feature set capturing how often the topic is changed in the group contributes to improve the accuracy for (IV) Optimism and (VI) Issue Involvement by +0.03, 0.01. From these results, each type of linguistic features contributes to improving the prediction accuracy of each type of performance index. In Section A, we performed a correlation analysis and discuss important indicators to capture highly performance.

8 CONCLUSION

This paper presented computational analysis for predicting the group performance which is defined by "product dimensions" [Hackman 1967 et al.] (PD). Group-level linguistic features are extracted including vocabulary, topic transition, and functional features. We also extracted nonverbal features from the speech turn, prosody, and head movement. With the MATRICS corpus with PD annotation, which includes three types of discussion tasks, we developed the discussion-task-dependent model for predicting PD of the unseen group which has discussed on the known task, and discussion-task-independent model for predicting PD of the unseen group which has discussed on the unknown task. In particular, the second prediction modeling was novel challenge. The experimental results show that a support vector regression model archived a 0.76 correlation in the discussion-task-dependent setting and 0.55 in the task-independent setting. Though multimodal fusion is effective for task-dependent prediction, it is not effective for task-independent prediction. The findings in this novel challenge leads two lines of future works. A direction of future work is to explore the common representation of multimodal features [3] to improve the accuracy for test data collected on unseen discussion task. Another direction is to explore the adaptation techniques using transfer learning [28] for the purpose.

ACKNOWLEDGMENT

This research is partially supported by KAKENHI: Grant-in-Aid for Scientific Research, Grant No. 19H01120, 19H01719.

REFERENCES

[1] Oya Aran and Daniel Gatica-Perez. 2013. One of a Kind: Inferring Personality Impressions in Meetings. In *Proceeding of ACM ICMI*. 11–18.

[2] Umut Avci and Oya Aran. 2016. Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. *IEEE Transaction on Multimedia* 18, 4 (2016), 643–658.

[3] T. Baltrušaitis, C. Ahuja, and L. Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (Feb 2019), 423–443.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[5] Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation* 41, 2 (2007), 181–190.

[6] Launor Carter, William Haythorn, and Margaret Howell. 1950. A further investigation of the criteria of leadership. *The Journal of Abnormal and Social Psychology* 45, 2 (1950), 350.

[7] Lei Chen, R. Travis Rose, Ying Qiao, Irene Kimbara, Fey Parrill, Haleema Welji, Tony Xu Han, Jilin Tu, Zhongqiang Huang, Mary Harper, Francis Quek, Yingen Xiong, David McNeill, Ronald Tuttle, and Thomas Huang. 2006. VACE Multimodal Meeting Corpus. In *Proceedings of MLMI*. 40–51.

[8] Mark G Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Proceedings of AAAI fall symposium on communicative action in humans and machines*, Vol. 56. Boston, MA, 28–35.

[9] Wen Dong, Taemie Kim, and Alex Pentland. 2009. A quantitative analysis of the collective creativity in playing 20-questions games. In *Proceedings of ACM conference on Creativity and cognition*. 365–366.

[10] Wen Dong and Alex "Sandy" Pentland. 2010. Quantifying Group Problem Solving with Stochastic Analysis. In *Proceedings of ICMI-MLMI*. Article 40, 4 pages.

[11] Daniel Gatica-Perez. 2009. Automatic nonverbal analysis of social interaction in small groups: A review. *Image Vision Computing* 27, 12 (nov 2009), 1775–1787.

[12] Daniel Gatica-Perez, Oya Aran, and Dinesh Babu Jayagopi. 2017. Analysis of Small Groups. *Chapter 25, Social Signal Processing* (2017), 349–367.

[13] J Richard Hackman, Lawrence E Jones, and Joseph E McGrath. 1967. A set of dimensions for describing the general properties of group-generated written passages. *Psychological Bulletin* 67, 6 (1967), 379.

[14] H. Hung and D. Gatica-Perez. 2010. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia* 12, 6 (Oct 2010), 563–575.

[15] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. I–364–I–367.

[16] Dinesh Babu Jayagopi, Dairazalia Sanchez-Cortes, Kazuhiro Otsuka, Junji Yamato, and Daniel Gatica-Perez. 2012. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proceeding of ACM ICMI*. 433–440.

[17] Taku KUDO. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of EMNLP*. 230–237.

[18] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[19] Patrick R Laughlin. 1980. Social combination processes of cooperative problem-solving groups on verbal intellective tasks. *Progress in social psychology* 1 (1980), 127–155.

[20] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceeding of ICML*. 1188–1196.

[21] Yun-Shao Lin and Chi-Chun Lee. 2018. Using Interlocutor-Modulated Attention BLSTM to Predict Personality Traits in Small Group Interaction. In *Proceedings of ACM ICMI*. 163–169.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[23] Michael T Mills and Nikolaos G Bourbakis. 2014. Graph-based methods for natural language processing and understanding—a survey and analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44, 1 (2014), 59–71.

[24] Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. (2018), 14–20.

[25] Fumio Nihei, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. 2014. Predicting Influential Statements in Group Discussions Using Speech and Head Motion Information. In *Proceedings of ACM ICMI*. 136–143.

[26] Fumio Nihei, Yukiko I Nakano, and Yutaka Takase. 2017. Predicting meeting extracts in group discussions using multimodal convolutional neural networks. In *Proceedings of ACM ICMI*. 421–425.

[27] Shogo Okada, Yoshihiko Ohtake, Yukiko I Nakano, Yuki Hayashi, Hung-Hsuan Huang, Yutaka Takase, and Katsumi Nitta. 2016. Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In *Proceedings of ACM ICMI*. 169–176.

[28] S. J. Pan and Q. Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (Oct 2010), 1345–1359.

[29] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. 2013. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces* 7, 1-2 (2013), 39–53.

[30] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. *IEEE Transaction on Multimedia* 14, 3-2 (2012).

[31] M. E. Shaw. 1973. Scaling group tasks: A method for dimensional analysis. *JSAS Catalog of Selected Documents in Psychology* 3, 8 (1973).

[32] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*. 97–100.

[33] Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

[34] Alexander Waibel and Rainer Stiefelhagen. 2009. *Computers in the Human Interaction Loop* (1st ed.). Springer Publishing Company, Incorporated.

[35] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

A CORRELATION ANALYSIS BETWEEN THE PRODUCT DIMENSIONS AND FEATURES

Table 6 shows the features that have the five highest Spearman's correlation coefficients r between the product dimensions and features in some features, where $p > 0.01, r > 0.468$. (+/-) denotes the sign (positive/negative) of the correlation. From table 6, acoustic features and linguistic features

Table 6: Features significantlly correlated to product dimensions (The listed features have a significant correlation where $p > 0.01, r > 0.468$. (+/-) denotes the sign (positive/negative) of the correlation. "DA" denotes dialog act bigram features)

Dimension	A, S	L
(I) Action Orientation	var F0 (+), var voiceProb (+) speaking time (max, total, mean) (+)	vec d44, d169, num of Noun (first use) total (+) DA(affirmation ⇄ suggestion) (+)
(III) Originality	var pcm zcr (+), var mfcc (-) speaking time (max-min (1s), std (1s)) (+)	vec d195, d38, d134, d61 (+) DA(suggestion → suggestion) (+)
(IV) Optimism	var F0 (+), var mfcc (+)	vec d88, d96, d180 (+) DA(understanding check ⇄ suggestion) (+)
(V) Quality of Presentation	var voiceProb (+), var mfcc (+), mean mfcc (+)	vec d169 (+), vec d138, d79 (-)
(VI) Issue Involvement	var pcm RMSenergy (-), mean mfcc (-), var mfcc (-) speaking time (max-min (1s), std (1s)) (+)	vec d44, d24, d38 (+) DA(suggestion → suggestion), DA(other → suggestion) (+)

significantly correlated with all performance aspects with the significant level. Conversely, any motion feature was not correlated with all aspects. For acoustic features, variance of F0 has positive correlation with (I) Action Orientation and (IV) Optimism. That of Zero-crossing rate (zcr) has positive correlation with (III) Originality. Features of mfcc has negative correlation with (III) Originality and (VI) Issue Involvement, but it has positive correlation with (IV) Optimism and (V) Quality of Presentation. For speaking turn features, speaking time features have positive correlation with (I), (III), (VI). For linguistic features, vocabulary features are highly correlated with all aspects. The interesting findings that the type of dialog act features which with significant correlation is different per aspect. The pairs with positive correlation of dialog act bi-gram features and aspects are "affirmation ⇄ suggestion " for (I), "suggestion → suggestion" for (III) and (VI), and "Understanding check ⇄ suggestion" for (IV). The consequent suggestions by multiple members ("suggestion → suggestion") seems to be a indicator to capture the highly performance. The "affirmation" and "Understanding check" are also important indicators to capture it.