Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5749–5758
Marseille, 20-25 June 2022

Dialogue Collection for Recording the Process of Building Common Ground in a Collaborative Task

Koh Mitsuda[1], Ryuichiro Higashinaka[1], Yuhei Oga[2*], Sen Yoshida[1]
[1]NTT Human Informatics Laboratories, NTT Corporation, Japan
[2]Graduate School of Systems and Information Engineering, University of Tsukuba, Japan
{koh.mitsuda.td,ryuichiro.higashinaka.tp,sen.yoshida.tu}@hco.ntt.co.jp
s2020716@s.tsukuba.ac.jp

Abstract

To develop a dialogue system that can build common ground with users, the process of building common ground through dialogue needs to be clarified. However, the studies on the process of building common ground have not been well conducted; much work has focused on finding the relationship between a dialogue in which users perform a collaborative task and its task performance represented by the final result of the task. In this study, to clarify the process of building common ground, we propose a data collection method for automatically recording the process of building common ground through a dialogue by using the intermediate result of a task. We collected 984 dialogues, and as a result of investigating the process of building common ground, we found that the process can be classified into several typical patterns and that conveying each worker's understanding through affirmation of a counterpart's utterances especially contributes to building common ground. In addition, toward dialogue systems that can build common ground, we conducted an automatic estimation of the degree of built common ground and found that its degree can be estimated quite accurately.

Keywords: dialogue systems, data collection, common ground, collaborative task

1. Introduction

Dialogue systems need to be able to build common ground with users accurately for them to successfully perform joint activities (Clark, 1996; Traum, 1994; Kopp and Kramer, 2021). To develop dialogue systems able to handle common ground, we believe the key is to reveal the process of building common ground.

In previous studies, the process of building common ground has been investigated by analyzing a dialogue between people who accomplish a collaborative task (Benotti and Blackburn, 2021; Chandu et al., 2021). These studies mainly focused on finding the relationship between a dialogue and the final result of the task (Anderson et al., 1991; Foster et al., 2008; He et al., 2017); however, the studies on the process of building common ground itself have not been well conducted. A few exceptions include the study by Udagawa and Aizawa (2020), who associated reference expressions with objects in a collaborative task to analyze the process of building common ground. Bara et al. (2021) recorded intermediate common ground by having each worker answer questions about his/her understanding of his/her partner's belief and behavior in real time during a collaborative task in a virtual space. These studies manually record common ground, which is costly.

In this paper, we look into the process of building common ground by automatically recording the intermediate result of the task as a proxy for the common ground being built. Recording the intermediate result of the task enables us to collect dialogues with common ground at low cost and investigate the information that the humans mutually believe at each step of dialogue.

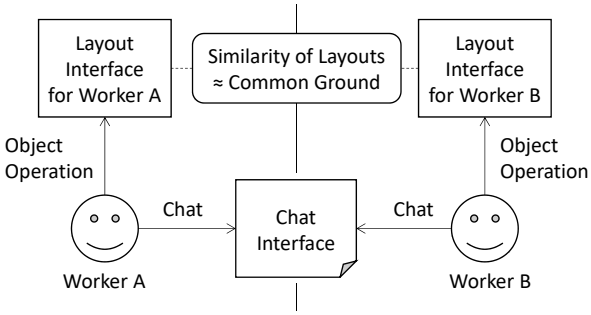

Figure 1: Proposed task: CommonLayout

We design a task called **CommonLayout** (Figure 1), in which two workers design the layout of objects into a common one through text chat. To quantify intermediate common ground through dialogue, we use the similarity of layouts created by two workers.

As a result of collecting such dialogues and investigating the process of building common ground, we found that the process can be classified into several typical clusters and that conveying each worker's understanding through affirmation of the counterpart's utterances especially contributes to building the common ground. In addition, toward the dialogue systems that can build common ground, we conducted an automatic estimation of the degree to which common ground is being built and found that its degree can be estimated quite accurately.

This paper makes three contributions:

- Propose a method to automatically and directly record the built common ground through dialogue. We can collect dialogues that show the process of

*Work carried out during internship at NTT.

building common ground without manual annotation.

- Clarify the process of building common ground. We found that the process of building common ground can be divided into typical clusters and that positive evaluations and empathy particularly contribute to building common ground.
- Conduct an experiment to automatically estimate the degree of common ground toward a dialogue system that can build common ground and show that the estimation is accurate.

Although our task may seem artificial, as far as we know, this study is the first attempt to quantify the process of building common ground. Since the process of building common ground can appear in any form of dialogue that needs mutual understanding, we consider our work to be a valuable contribution to dialogue system research.

2. Related Work

In previous studies, the process of building common ground has been investigated by analyzing human-human dialogue in which a collaborative task is performed (Benotti and Blackburn, 2021; Chandu et al., 2021). Common ground is considered to be successfully built when the task is successfully performed on the basis of the final result of the task. For example, the collaborative tasks for operating certain objects in maps (Anderson et al., 1991; Carletta et al., 1991; Bard et al., 2000; Denis and Striegnitz, 2012) and puzzles (Foster et al., 2008; Spanger et al., 2009; Tokunaga et al., 2012) have been investigated.

Recently, in the task of finding or creating certain objects, various collaborative games have been proposed regarding graphics (Liu et al., 2012; de Vries et al., 2017; Kim et al., 2019; Udagawa and Aizawa, 2019), virtual environments (Polyak and Davier, 2017; Ilinykh et al., 2019; Hahn et al., 2020; Jayannavar et al., 2020), and texts (He et al., 2017; Lewis et al., 2017; Gero et al., 2020). Another major line of studies, albeit in human-system dialogue, aims at manipulating robots in real space by generating utterances that can be grounded to physical objects (Moratz and Tenbrink, 2006; Hough and Schlangen, 2017; Chai et al., 2017; Van Waveren et al., 2019). In these studies, the relationship between the dialogue and common ground is analyzed on the basis of the final result of the task. This is because these studies aim to clarify what dialogue phenomena occur in the dialogue where the interlocutors need to build common ground for completing a task successfully. However, in such a framework, the process of building common ground is not recorded and thus cannot be analyzed.

Our research attempts to record the process of building common ground and is most closely related to the task of OneCommon (Udagawa and Aizawa, 2019; Udagawa and Aizawa, 2020) and MindCraft (Bara et al., 2021). In OneCommon (Udagawa and Aizawa, 2019), a graphical plane is prepared including multiple dots in random positions, sizes, and colors. Two workers are given only a slightly different portion of the plane, and each worker selects one dot through dialogue. The task is successful if the selected dots are identical between the workers. Grounding reference expressions to the dots between workers is considered to reflect the process of building common ground, and such expressions are manually annotated (Udagawa and Aizawa, 2020). In MindCraft (Bara et al., 2021), two players of Minecraft are given the knowledge and skills to work together to create a specific object. Every 75 seconds while proceeding with the task, a player must answer a predetermined question about the common ground (e.g., "What do you think your counterpart is building right now?"). Our research is similar to these studies but differs in two ways. First, we automatically record the common ground, making data collection less costly. Second, we clarify the process of building common ground on the basis of quantified common ground using the similarity of layouts.

3. Data Collection

In this section, after explaining the requirements of the task for recording the process of building common ground, we describe the task description and the process of data collection.

3.1. Requirements

For designing the task to reveal the process of building common ground, we followed OneCommon (Udagawa and Aizawa, 2019). OneCommon was proposed on the basis of the requirements that necessitated *continuous* and *partially-observable* contexts for introducing the difficulty of building common ground. The *continuous* context means that a task should include not only categorical information but also continuous information (thus, it is difficult to use symbolic expressions to describe the information) such as graphics, which makes it necessary for speakers to exchange multiple utterances to build common ground. The *partially-observable* context means that only partial information is shared among the workers. In such a context, the workers need to build common ground by sharing their own information through exchanges of utterances.

In addition to the requirements asserted in OneCommon, we defined two other requirements for our task. First, the task should require workers to perform multiple operations to complete it. This is because we want to collect intermediate task results as common ground. Second, the intermediate results of the task should be quantifiable in terms of its progress to quantify common ground being built.

3.2. Task Description

Figure 1 illustrates the proposed task, CommonLayout, in which two workers lay out the same figure set into a

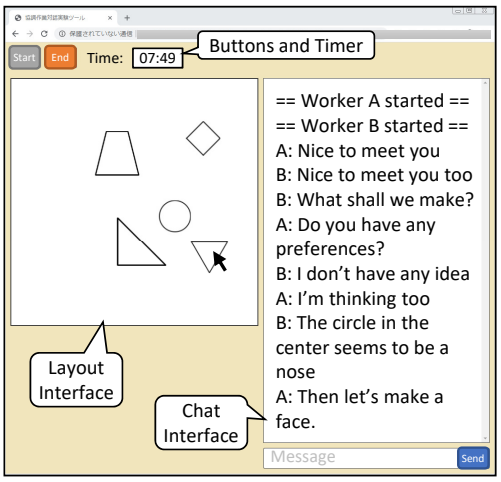

Figure 2: Interface for collecting dialogues in CommonLayout. This example shows the interface presented to one of the workers. Each worker can see the common chat interface and his/her layout interface but cannot see the partner's one. Dialogue shown was originally in Japanese and translated by authors.

common design through text chat. To perform the task, they discuss the idea of a final layout and move figures into the same position one by one. Note that each worker can see only his/her layout. To accomplish this task, the worker needs to imagine the partner's layout and move his/her figures. They need to discuss a policy of the final layout and move each figure step by step.

Figure 2 shows the interface for collecting dialogues in CommonLayout. The interface consists of a "layout interface" and "chat interface" with a "Start" button, "End" button, and countdown timer shown in the upper left. Objects in a randomized layout for each worker are shown on the layout interface at the beginning of the task. Workers are allowed to move a figure but not delete, rotate, or scale it. Dragging and dropping for moving figures are recorded as an operation log with coordinates and timestamps. The chat interface is visible to both workers, but each worker's layout interface is visible only to him/her. The end button is pressed if the workers believe that their layouts are identical. The task is finished when the end button is pressed by the two workers or the dialogue exceeds 10 minutes.

Figure 3 shows the prepared figure sets: **Simple** and **Building**. Simple consists of 10 basic figures (e.g., squares and circles), and Building consists of 10 building icons[1] (e.g., bar and police). With Building, workers can use the knowledge related to buildings, for example, "The police should be placed near the bar for security."

3.3. Collected Data

In the data collection, 287 workers familiar with PC operations participated. They were recruited through a recruitment agency and paid for their participation. The workers were allowed to quit the experiment at any time. Two workers who had never met were randomly put into pairs for a total of 213 pairs, and each pair conducted the task four times (five figures from Simple/Building and seven figures from Simple/Building in a randomized order). We instructed the workers to be creative in their layout and decide on the final layout with as much input from their partner as possible. The dialogues were conducted in Japanese.

Table 1 shows the statistics of the collected data. A total of 984[2] dialogues were collected. Each dialogue included 28.8 utterances on average. Building had more utterances per dialogue and fewer operations (moves) per dialogue than Simple. This may be because discussing the final idea is easier in Building with the knowledge of buildings than in Simple. Note that 288 pairs completed the task in less than 10 minutes with an average of 521 seconds, and the others worked for up to 10 minutes or until time ran out.

Table 2 shows a collected dialogue of CommonLayout. This example is the dialogue collected in the session shown in Figure 2. The idea of the final layout was agreed upon by U_{14}, and the figures are placed after U_{15}. In this dialogue, we can see the process of building common ground, such as the utterances from U_7 to U_{14}, in which the idea of the figure placement was decided through consultation, and U_{24} or U_{29}, in which the workers confirmed their placement.

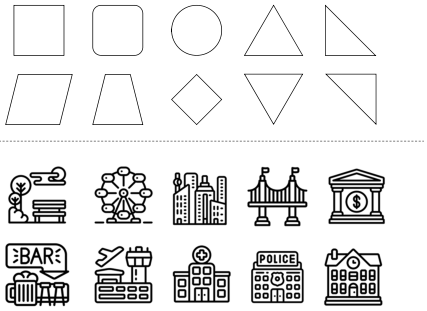

Figure 3: Figure sets Simple (upper) and Building (lower) used in CommonLayout

Figure set	Simple		Building		Total
No. of figures	5	7	5	7	
Workers	283	287	281	286	287
Worker pairs	210	213	209	212	213
Dialogues	245	244	248	247	984
Utterances	6,652	6,632	7,416	7,674	28,374
Utts./dialogue	27.1	27.1	29.9	31.0	28.8
Letters/utts.	14.5	14.3	14.7	14.1	14.4
Moves	14,580	19,584	12,700	17,286	64,420
Moves/dialogue	59.5	81.3	51.2	69.9	65.4

Table 1: Statistics of collected data

[1] https://www.flaticon.com/packs/city-life-3.

[2] Due to the difficulty of assigning workers, 24 of 213 pairs performed four sessions multiple times.

ID	S	Utterance
U_1	A	Nice to meet you.
U_2	B	Nice to meet you too.
U_3	B	What shall we make?
U_4	A	Do you have any preferences?
U_5	B	I don't have any idea.
U_6	A	I'm thinking too.
U_7	B	The circle in the center seems to be a nose.
U_8	A	Then, let's make a face.
U_9	A	The circle looks like a nose to me.
U_{10}	B	I agree with you.
U_{11}	B	How about Pinocchio?
U_{12}	A	That's good.
U_{13}	A	Shall we make Pinocchio?
U_{14}	B	Okay.
U_{15}	B	It moves, doesn't it?
U_{16}	A	I just tried to change the angle of the bottom triangle, but I can't seem to do it.
U_{17}	A	Let's create a tilted face.
U_{18}	B	That's good.
U_{19}	B	Let's move the mouth a little to the left.
U_{20}	A	What do you think?
U_{21}	A	I've just moved it.
U_{22}	B	That's good.
U_{23}	B	Put the nose in the middle a little bit.
U_{24}	A	What about the shape on the left?
U_{25}	B	I tried to make a downward-pointing triangle and a square for the eyes.
U_{26}	B	A trapezoid looks like a head forelock.
U_{27}	A	I see.
U_{28}	A	I've moved it around.
U_{29}	B	It looks like a face, doesn't it?
U_{30}	A	Yes, I think so too.
U_{31}	A	I think it should look like this.
U_{33}	B	It looks good.

Table 2: Example of collected dialogue in Common-Layout. ID and S corresponds to utterance ID and speaker. This dialogue is collected in the session shown in Figure 2. Utterances were translated from Japanese by the authors.

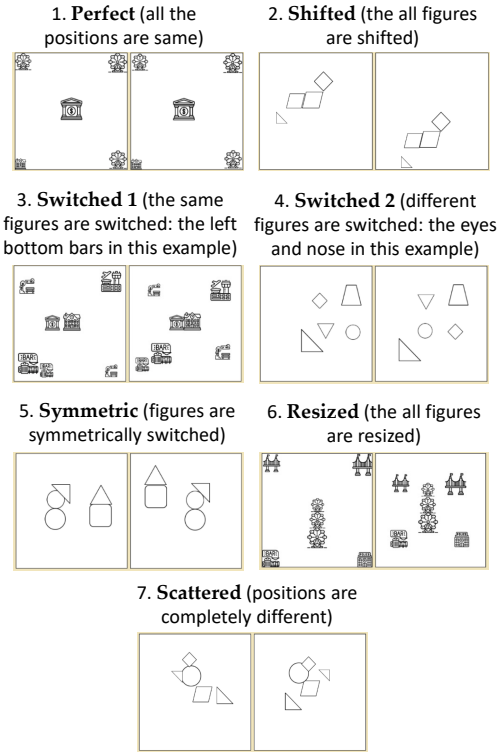

Figure 4: Final layout patterns

	Figure set	Simple		Building		Total
	No. of figures	5	7	5	7	
1	Perfect	13%	13%	26%	22%	19%
2	Shifted	15%	5%	5%	4%	7%
3	Switched 1	0%	1%	3%	4%	7%
4	Switched 2	32%	45%	31%	40%	37%
5	Symmetric	3%	3%	3%	4%	3%
6	Resized	10%	5%	19%	12%	12%
7	Scattered	27%	27%	13%	15%	20%

Table 3: Frequency of final layout patterns

4. Analysis

With the collected data, we analyzed the process of building common ground. First, we manually checked whether the final results of layout pairs were identical. We then quantified the intermediate task results as common ground on the basis of the similarity of the layouts. Finally, we applied time-series analysis to the quantified common ground and found typical clusters of building common ground.

4.1. Patterns of Final Common Ground

From the collected data, we found that certain pairs of the final layouts were not identical even though most workers judged the task as accomplished. To quantify common ground, we first identify patterns in the pairs of the final layouts and then determine a measure on the basis of the patterns of success and failure.

Figure 4 shows the final layout patterns. We manually checked 50 randomly selected pairs of the final layouts. We found that final layout patterns can be categorized into seven patterns. The patterns, except for Perfect, had different positions of figures. In addition to Perfect, we decided to regard Shifted as success of the task. This is because we did not instruct the workers to put the figures into the same position in absolute coordinates.

Table 3 shows the frequency of final layout patterns in all samples annotated by a worker different from the authors. The table shows that the success rate (Perfect or Shifted) was 25% (= 19% + 6%). Regarding the figure sets, the success rate was higher in Building than in Simple. Regarding the number of figures, the frequency of each pattern was similar between five and seven; thus, the difficulty of the task does not depend on the number of figures. Interestingly, there were many cases in which the users failed in the task, making the data a valuable resource for analyzing how common ground can/cannot be built within a dialogue.

4.2. Quantification of Common Ground

On the basis of the final layout patterns, we introduce a measure to quantify common ground. We assume that the task succeeds at pattern 1–2 (Perfect and Shifted), almost succeeds at pattern 3–6 (Switched 1, Switched 2, Symmetric, and Resized), and fails at pattern 7 (Scattered). We call these patterns **Success**, **Middle**, and **Fail**. We introduce **layout distance**, which becomes small in Success. This measure is the sum of the distances between two arbitrary figures in a layout pair, as shown in the following equation.

$$distance(L_A, L_B) = \sum_{i,j \in Figures} \|\vec{a}_{i,j} - \vec{b}_{i,j}\|,$$

L_A and L_B are the layouts created by workers A and B, and $\vec{a}_{i,j}$ and $\vec{b}_{i,j}$ are the vectors (representing x and y coordinates) defined between figures i and j in L_A and L_B. The smaller the Euclidean distance, the more common ground is considered to have been built. On the basis of the layout distance, we can quantify the degree of common ground in each time step within a dialogue.

Figure 5 shows the degree of common ground in each time step[3] on average. Note that an utterance is regarded as one time step, and moving the same figure multiple times in succession is also regarded as one time step. The layout distance at the end of the task (when the time step is 50) is larger in the order of Success, Middle, and Fail, confirming that the degree of common ground can be quantified as we defined it.

To reveal the typical process of building common ground, we used k-Shape, which is a time-series clustering method based on k-means (Paparrizos and Gravano, 2015). We applied k-Shape to all 987 dialogues (i.e., 987 trajectories of layout distances) in the collected data. Figure 6 shows the results of time-series clustering on the layout distance. We set the number of clusters to five because similar clusters appear if this number is increased, and fewer k led to clusters with inconspicuous traits. The x-axis shows time steps, and the y-axis shows the z-normalized layout distance. The ratios of the number of dialogues for the five clusters (Clusters 1 to 5) were 29%, 28%, 24%, 12%, and 7%, respectively.

In Cluster 1, the layout distance consistently decreased throughout the dialogue; thus, common ground seems to have been smoothly built. From Clusters 2 to 4, building common ground stagnated. In these clusters, the workers discussed the positions of partial figures until the middle of the dialogue. They then confirmed the idea of a final layout and moved the remaining figures to the same positions towards the end. Cluster 5 corresponds to Scattered in final layout patterns where common ground was not properly built.

[3]We normalized the total steps in each dialogue to 50 steps by linear transformation.

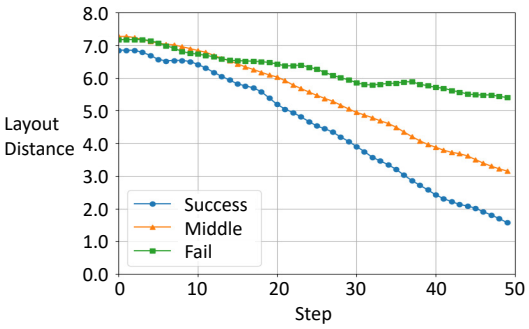

Figure 5: Degree of common ground in each time step on average. Success, Middle, and Fail correspond to patterns 1–2, 3–6, and 7 in Figure 4.

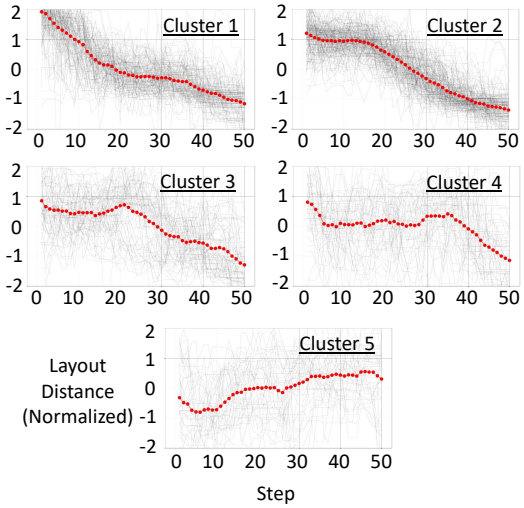

Figure 6: Clusters representing process of building common ground

Results	C1	C2	C3	C4	C5
Success	28%	32%	27%	15%	10%
Middle	58%	57%	58%	51%	46%
Fail	14%	11%	15%	34%	44%

Table 4: Correspondence between task results and clusters (denoted as C1–C5) in the process of building common ground shown in Figure 6.

Table 4 shows the correspondence between the task results and the clusters in the process of building common ground. Each cell in the table shows the ratio of task results in each cluster. Clusters 1 to 3 contain many patterns corresponding to Success and Middle, indicating that common ground was successfully built. Clusters 4 and 5 contain many Scattered corresponding to Fail, indicating a process in which the common ground was not successfully built. From these results, in the case of Success, common ground starts being built relatively early in the process and becomes sufficient by the end of the dialogue. In the case of Fail, the common ground starts to be built from the latter part of the dialogue and is not sufficient by the end of the task, or the common ground cannot be built through the whole dialogue.

	Success	Middle	Fail
1	Good**	I did it**	What do we do?**
2	That's good**	I placed it**	Small**
3	Is it good?**	I put it*	Well**
4	Let's do it**	I agree*	Let's use it**
5	It's done**	Good*	I'm looking at it**

Table 5: Top five statistically significant occurrences of linguistic expressions in dialogues divided by the results of the task. "**" and '*' denotes $p<.01$ and $p<.05$ in statistical test.

	Success	Middle	Fail
1	Self-disclosure: preference (+)*	Sympathy**	Information**
2	Thanks*	Question: Preference*	Greeting*
3	Sympathy*	–	Self-disclosure: preference (–)*

Table 6: Top three statistically significant occurrences of dialogue acts in dialogues divided by the results of the task. Symbols of '+' and '–' indicate positive and negative in "self-disclosure: preference."

4.3. Process of Building Common Ground

On the basis of final layout patterns and layout distance, we look into linguistic phenomena in the dialogues where common ground is successfully built or not. These phenomena are important to create a dialogue system that can build common ground with users. We first divided dialogues into three types (Success, Middle, and Fail) and investigated the linguistic expressions and dialogue acts that frequently occur in Success, Middle, or Fail. In addition, we used a Hidden Markov Model (HMM) to model the transitions of dialogue acts in each kind of dialogue and then investigated which transitions are especially effective to build common ground.

Table 5 shows the top five statistically significant occurrences of linguistic expressions in the dialogues divided by the results of the task (Success, Middle, and Fail). Each expression represents a predicate (translated by the authors). A Japanese morphological analyzer JTAG (Fuchi and Takagi, 1998) was used to extract the expressions. Fisher's exact probability test was used as the statistical test, and the expressions that appeared significantly more often in dialogues categorized in one of the three types were obtained by sorting them in ascending order by p-value. From this table, regarding Success, expressions that indicate positive evaluation such as "Good" and those that indicate graphic manipulation such as "Let's do it" and "It's done" appear more frequently. As for Middle, instead of positive expressions, there are many expressions showing sympathy such as "I agree" and graphic manipulations such as "I did it." In the case of Fail, the expressions regarding positive evaluations and graphic manipulations do not appear, and ambiguous expressions such as "What do we do?" and "I'm looking at it" appeared more frequently. These results suggest that evaluation expressions or sympathetic expressions that convey one's understanding and behavior are important in building common ground successfully.

Table 6 shows the top three statistically significant dialogue acts that appear in the dialogue divided by the results of the task. We used the label set of dialogue acts proposed by Meguro et al. (2011). There are 33 types for dialogue acts (e.g., self-disclosure: disclosure of preferences and feelings, information: delivery of objective information, and sympathy: sympathetic utterances and praise). For the estimation, we used the support vector machine trained by (Higashinaka et al., 2014). As in the analysis for Table 5, Fisher's exact probability test was used for listing the dialogue acts. From this table, we can see that in Success, there are many expressions that affirm the counterpart, such as self-disclosure of positive preference, thanks, and empathy. Similar to Success, many expressions of empathy appeared in Middle. There are also many questions about the evaluation. In the case of Fail, there are many expressions that are different from those of Success and Middle, such as information and greetings. Expressions related to evaluation also appeared, but their contents were negative. These results suggest that, as in the case of Table 5, self-disclosure regarding positive evaluation and empathy that conveys one's understanding as well as information that conveys one's behavior is important for building common ground.

The analysis up to here has focused on linguistic expressions and dialogue acts; by focusing on the transitions of the dialogue act and layout distance, it may be possible to clarify what kind of dialogue particularly contributes to the process of building common ground. To investigate the transition of the dialogue acts, we modeled it using a HMM, which is commonly used to learn the structure of data series with an unknown number of states (Rabiner and Juang, 1986). The HMM library, `hmmlearn`[4], was used for modeling. The learning method followed that of (Meguro et al., 2014). Specifically, we initialized the ergodic HMM so that only the dialogue acts of a specific speaker (Speaker A) were output from half of the states, and only the dialogue acts of the counterpart speaker (Speaker B) were output from the other half of the states. The number of states ranged from 1 to 10, and 10 HMMs were trained in each number of states, resulting in 100 HMMs. The optimal HMM was selected using the Minimum Description Length (MDL) criterion.

Figures 7 and 8 show the HMMs constructed from the series of dialogue acts in Success and Fail. One of Middle is omitted because a similar HMM was constructed for Success. In each state, 'A' and 'B' represent the two speakers, and dialogue acts with observation probability are listed. 'p' on the edges represents the transition

[4] https://github.com/hmmlearn/hmmlearn

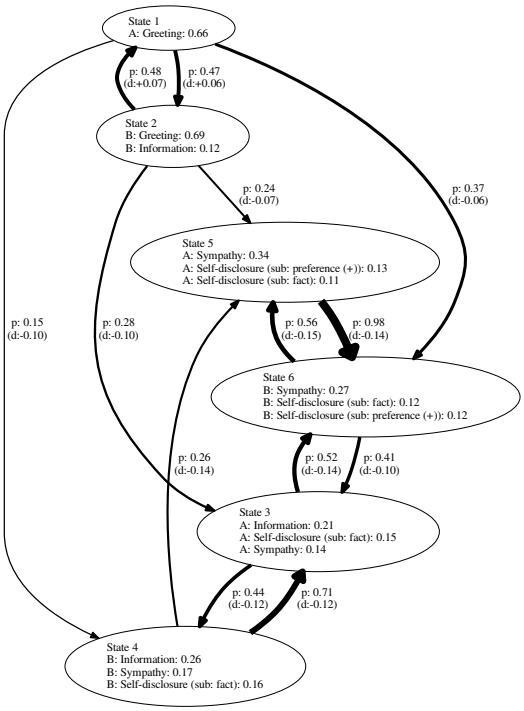

Figure 7: HMM created from a series of dialogue acts in dialogues categorized in Success. 'A' and 'B' represent two speakers, and 'p' and 'd' on edge represent transition probability and average difference of layout distance in transition. Edge is thicker when transition probability is higher.

probability and corresponds to thickness of the edges. 'd' on the edges represent the average difference in layout distance on the transitions. It was calculated by weighting the average difference of layout distance for each dialogue act pair included in two adjacent states by the observation probability of that pair. A negative value for the difference in layout distance on the transition indicates the progress of building common ground.

From these two HMMs, these dialogues are completely different and indicate that the dialogues in Success are more complicated than those in Fail. We can see that the Success and Fail dialogues contain both common and different structures. As common structures, we see a pair of initial greetings (states 1 and 2 in both HMMs) and a pair of information or empathy (states 3 and 4 in both HMMs). We also see a different structure, a pair involving empathy and self-disclosure of positive preference (states 5 and 6 in the HMM of Figure 7). This structure has the transition with the smallest value of difference (-0.15) for layout distance, indicating that the common ground is particularly well developed in this transition period. This result suggests that it is most important to check each other's understanding through self-disclosure regarding evaluation and sympathy to build common ground.

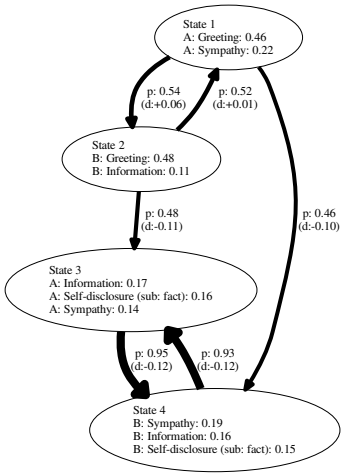

Figure 8: HMM created from a series of dialogue acts in dialogues categorized in Fail

5. Estimation Presentments

Toward dialogue systems that can build common ground with users, we conducted an automatic estimation of the degree of built common ground. We address the problem of estimating the degree to which common ground is being built, represented by layout distance. The definition of the problem is that, at a certain time step i in the dialogue, given the dialogue context (the utterances from U_1 to U_i) and layout $L_{X,i}$, where X is a speaker of U_i, the layout distance at the time step i is estimated without the information about the layout of worker X's partner.

5.1. Model and Data Preparation

Figure 9 shows the model architecture for estimating the layout distance from the dialogue and one's own layout. We trained a model as in Soleymani et al. (2019), which takes into account both the dialogue and layout by using pre-trained models. We used Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) for the dialogue and Residual Network (ResNet) (He et al., 2016) for the layout as pre-trained models. The embeddings obtained from the pre-trained models were concatenated and converted into a vector with fully connected layers, finally producing a scalar value denoting the layout distance (i.e., the degree of common ground).

The collected 984 dialogues were divided into 8:1:1 (training, development, and test set). We split the data in a manner in which the dialogues by the same pair of workers were not included in different sets. The layout distance was normalized from zero to one using min-max normalization. The pre-trained models were a BERT-base model trained for Japanese from huggingface/transformers[5] and ResNet-18 model from

[5] https://huggingface.co/cl-tohoku

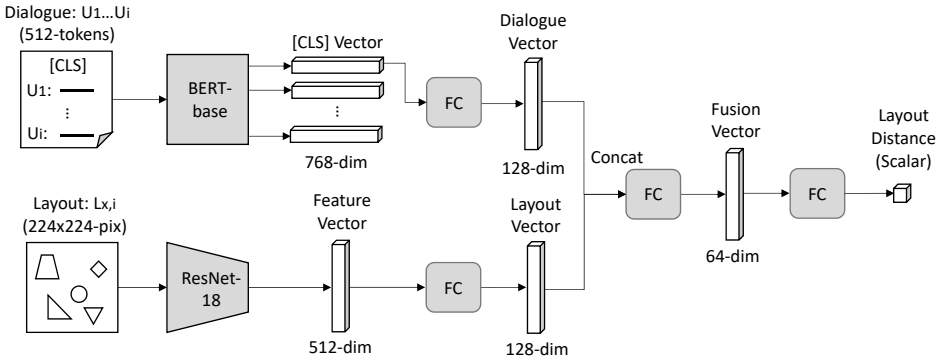

Figure 9: Model architecture for estimating layout distance from given dialogue and layout (Dialogue+Layout). Layout corresponds to layout of speaker X when X utters U_i. FC denotes fully connected layer.

torchvision[6]. For fine-tuning the models, we used mean squared error (MSE) as a loss function and Adam (Kingma and Ba, 2015) as an optimizer, where the learning rate was set to 2e-5.

Four types of models were prepared for investigating the effectiveness of the information used in training. "Mean-Baseline" is a baseline that outputs a constant value (the average value in the training set). "Dialogue" and "Layout" are models taking into account only one of the dialogues or layouts as input. "Dialogue+Layout" is a model shown in Figure 9 that takes both into account.

Model	MSE
(a) Mean-Baseline	0.0274
(b) Dialogue	0.0226_a
(c) Layout	0.0225_a
(d) Dialogue+Layout	$\mathbf{0.0178}_{abc}$

Table 7: Mean squared errors (MSEs) between estimated and true layout distance from given dialogue and layout. Subscripts indicate that a score is significantly better ($p<.05$) than those of corresponding models in Steel-Dwass test.

5.2. Results

Table 7 shows the MSEs between the estimated and true layout distance from the given dialogue and layout. The MSE for Dialogue+Layout was the lowest with significant differences ($p<.05$) in a Steel-Dwass multiple comparison test (Dwass, 1960), confirming the effectiveness of considering both dialogue and layout. The performances of Dialogue and Layout were lower than that of Dialogue+Layout. This may be because Dialogue+Layout determined which figures were grounded from the layout information with the content of the dialogue. Since the MSE for Dialogue+Layout was 0.0178, the error between the estimated layout distance and true value was about +/– 0.134 on average, which confirms that the estimation is quite accurate. These results suggest that it will be possible to create a dialogue system that can perform CommonLayout with users while understanding the degree of built common ground.

6. Conclusion

To reveal the process of building common ground in dialogue, we devised a task called CommonLayout, in which two workers collaboratively placed figures in a common layout. For automatically recording the process of building common ground through a dialogue, we utilized intermediate task results where commonality of layouts between the workers is regarded as common ground being built within a dialogue. We collected 984 dialogues where workers performed CommonLayout. By investigating the collected data, we found seven final layout patterns. On the basis of these final layout patterns, we introduced layout distance as quantification of common ground. Time-series clustering was applied to the transitions of the layout distance for revealing a typical process of building common ground. We found that the process of building common ground can be divided into five typical clusters. In addition, we analyzed the linguistic phenomena that lead to task accomplishment in terms of dialogue and layout distance. The results suggest that conveying one's understanding to others through positive evaluation and empathy is the most important factor in building common ground. We also found that the degree of common ground being built can be estimated to some extent.

Future work includes automatically estimating a partner's figure layout from dialogue and one's own layout since dialogue systems must be able to understand the belief of a partner for better task success. Eventually, we want to integrate such an estimator into a dialogue system so that it can successfully perform CommonLayout with users. For this purpose, we can refer to the dialogue model proposed by Fried et al. (2021), which successfully accomplished OneCommon with users. In addition, a method needs to be developed for quantifying common ground in other collaborative tasks, such as those requiring more elaborate and complex interactions, to verify the generalizability of the results.

[6] https://pytorch.org/vision/stable/models.html

7. References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC map task corpus. *Computational Linguistics*, 34(4):351–366.

Bara, C.-P., CH-Wang, S., and Chai, J. (2021). MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proc. of EMNLP*, pages 1112–1125.

Bard, E. G., Anderson, A. H., Sotillo, C., Doherty-Sneddon, M. A. G., and Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, 42:1–22.

Benotti, L. and Blackburn, P. (2021). Grounding as a collaborative process. In *Proc. of EACL*, pages 515–531.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., and Anderson, A. H. (1991). The reliability of dialogue structure coding scheme. *Language and Speech*, 23(1):13–32.

Chai, J. Y., Fang, R., Liu, C., and She, L. (2017). Collaborative language grounding toward situated human-robot dialogue. *AI Magazine*, 37(4):32–45.

Chandu, K. R., Bisk, Y., and Black, A. W. (2021). Grounding 'grounding' in NLP. In *Findings of ACL-IJCNLP*, pages 4283–4305.

Clark, H. H. (1996). *Using language*. Cambridge university press.

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., and Courville, A. C. (2017). GuessWhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*, pages 5503–5512.

Denis, A. and Striegnitz, K. (2012). A collaborative puzzle game to study situated dialog. In *Proc. of AAAI*, pages 37–40.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.

Dwass, M. (1960). Some k-sample rank-order tests. *Contributions to probability and statistics*, pages 198–202.

Foster, M. E., Bard, E. G., Guhe, M., Hill, R. L., Oberlander, J., and Knoll, A. (2008). The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proc. of HRI*, pages 295–302.

Fried, D., Chiu, J., and Klein, D. (2021). Reference-centric models for grounded collaborative dialogue. In *Proc. of EMNLP*, pages 2130–2147.

Fuchi, T. and Takagi, S. (1998). Japanese morphological analyzer using word co-occurrence -JTAG-. In *Proc. of COLING*, pages 409–413.

Gero, K. I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruiz, M., Miller, S., Millen, D. R., Campbell, M., Kumaravel, S., and Zhang, W. (2020). Mental models of AI agents in a cooperative game setting. In *Proc. of CHI*, pages 1–12.

Hahn, M., Krantz, J., Batra, D., Parikh, D., Rehg, J. M., Lee, S., and Anderson, P. (2020). Where are you? localization from embodied dialog. In *Proc. of EMNLP*, pages 806–822.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. of CVPR*, pages 770–778.

He, H., Balakrishnan, A., Eric, M., and Liang, P. (2017). Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proc. of ACL*, pages 1766–1776.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (2014). Towards an open domain conversational system fully based on natural language processing. In *Proc. of COLING*, pages 928–939.

Hough, J. and Schlangen, D. (2017). It's not what you do, it's how you do it: Grounding uncertainty for a simple robot. In *Proc. of HRI*, pages 1–10.

Ilinykh, N., Zarrieß, S., and Schlangen, D. (2019). Meet Up! a corpus of joint activity dialogues in a visual environment. In *Proc. of SEMDIAL*, pages 1–10.

Jayannavar, P., Narayan-Chen, A., and Hockenmaier, J. (2020). Learning to execute instructions in a Minecraft dialogue. In *Proc. of ACL*, pages 2589–2602.

Kim, J.-H., Kitaev, N., Chen, X., Rohrbach, M., Zhang, B.-T., Tian, Y., Batra, D., and Parikh, D. (2019). CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proc. of ACL*, pages 6495–6513.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. of ICLR*, pages 1–15.

Kopp, S. and Kramer, N. (2021). Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, 12.

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., and Batra, D. (2017). Deal or no deal? end-to-end learning for negotiation dialogues. In *Proc. of EMNLP*, pages 2443–2453.

Liu, C., Fang, R., and Chai, J. Y. (2012). Towards mediating shared perceptual basis in situated dialogue. In *Proc. of SIGDIAL*, pages 140–149.

Meguro, T., Higashinaka, R., Minami, Y., and Dohsaka, K. (2011). Evaluation of listening-oriented dialogue control rules based on the analysis of HMMs. In *Proc. of Interspeech*, pages 809–812.

Meguro, T., Minami, Y., Higashinaka, R., and Dohsaka, K. (2014). Learning to control listening-oriented dialogue using partially observable markov

decision processes. *ACM Transactions on Speech and Language Processing*, 10(4).

Moratz, R. and Tenbrink, T. (2006). Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition & Computation*, 6(1):63–107.

Paparrizos, J. and Gravano, L. (2015). k-Shape: Efficient and accurate clustering of time series. In *Proc. of SIGMOD*, pages 1855–1870.

Polyak, S. and Davier, A. V. (2017). Analyzing game-based collaborative problem solving with computational psychometrics. In *Proc. of SIGKDD*, pages 1–14.

Rabiner, L. and Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP magazine*, 3(1):4–16.

Soleymani, M., Stefanov, K., Kang, S.-H., Ondras, J., and Gratch, J. (2019). Multimodal analysis and estimation of intimate self-disclosure. In *Proc. of ICMI*, pages 59–68.

Spanger, P., Yasuhara, M., Iida, R., and Tokunaga, T. (2009). Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task. In *Proc. of preCogsci*, pages 1–8.

Tokunaga, T., Iida, R., Terai, A., and Kuriyama, N. (2012). The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proc. of LREC*, pages 422–429.

Traum, D. R. (1994). A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science.

Udagawa, T. and Aizawa, A. (2019). A natural language corpus of common grounding under continuous and partially-observable context. In *Proc. of AAAI*, pages 7120–7127.

Udagawa, T. and Aizawa, A. (2020). An annotated corpus of reference resolution for interpreting common grounding. In *Proc. of AAAI*, pages 9081–9089.

Van Waveren, S., Carter, E. J., and Leite, I. (2019). Take one for the team: The effects of error severity in collaborative tasks with social robots. In *Proc. of IVA*, pages 151–158.