Construction of Multimodal Chat-talk Corpus Considering the Closeness in Dyads and Analysis of Dialog Act

Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito
Graduate School of Engineering,
Tohoku University, Japan
Email: {yoshihiro.yamazaki.t2@dc, yuya@spcom.ecei, tnose@m, aito@spcom.ecei}.tohoku.ac.jp

***Abstract*—Multimodal dialog systems that can make small talks with facial expression, gesture, and gaze are expected as a next-generation dialog-based system. Such dialog systems should consider a social relationship with a dialog partner to be accepted by human-society as a social existence. However, a sufficient amount of a multimodal dialog data corpus considering the closeness between the speakers has never constructed so far. Therefore, it is still not fully understood how the system should control the dialog behavior according to closeness to the user. In this study, we constructed a multimodal Japanese chat-talk corpus and analyze the dialog behaviors toward the modeling of the dialog strategy considering the closeness to the user. The constructed dialog corpus contains 19,303 utterances (10 hours) from 19 pairs of participants. The analysis shows that closeness affects dialog acts, and we obtain clues to model the dialog strategy considering the relationship with the user.**

***Keywords*—interpersonal closeness; multimodal interaction; chat-talk corpus; dialog act**

I. Introduction

In recent years, spoken dialog systems are widely used as a smart speaker and a communication robot. Many of such systems have a function that can entertain their users by making chat-like talks. To achieve natural short talks, response generation based on deep learning has been studied actively (e.g., [1]). Such neural conversational models are usually trained using a large-scale text-based conversational dataset, such as chats on Twitter [2]. On the other hand, non-verbal information is also important for the dialog system to achieve natural conversation. Actually, many studies have focused on the multimodal dialog systems [3], [4]. To decide the optimum dialog strategy of multimodal dialog systems, a sufficient amount of the conversational data which contains the non-verbal information, such as facial expression, gaze, and gesture is required. Some studies constructed the dialog corpora contain the multimodal information [5], but it is still not enough to construct the dialog system, particularly regarding Japanese dialog corpora.

On the other hand, an open-domain dialog system has been attracted attention. The chat-oriented dialog systems are expected to entertain their users, and various studies that focus on user evaluation have been conducted. For example, some studies revealed that changing dialog responses depending on the closeness to the user improves user impression [6], [7]. Although these studies have controlled a form of an utterance, other dialog behavior should change depending on the closeness in dyads. According to the social penetration theory [8], individuals become more comfortable to talk about private and personal topics (i.e., self-disclosure) as relationships become closer in human-human conversation. Therefore, the dialog systems might further improve the user evaluation if they can control the utterance at the intention level based on the relationship with the user.

In this study, we construct a large scale multimodal Japanese chat-talk corpus with the information of the closeness between speakers. Audio signals and video clips of the collected corpus are clear enough to use for the research of the multimodal open-domain dialog system. In addition, the corpus is also useful to construct the statistical dialog model considering the closeness to the dialog partner. In this paper, we focus on the intention of the utterance and compare the frequency of the dialog acts and transition of them between different closeness levels.

II. Conventional corpora

Some studies tried to construct the multimodal dialog corpus so far. For example, Santa Barbara Corpus of Spoken American English (SBCSAE) [9] and The Ritel Corpus [10] are typical corpora. Table I shows the comparison between these two corpora and our target corpus. SBCSAE collected a wide variety of spontaneous conversations in everyday lives. This corpus contains rich interactions of daily human-human conversations, but the structure of the collected dialog is complex because the dialog was conducted by multi-persons. Many of the current dialog systems assume a single dialog partner as the user, and it is difficult to correspond to the multi-party dialog. In addition, the audio signal of the corpus has many overlaps of speech, and it is not easy to conduct the analysis. On the other hand, The Ritel Corpus contains the human-machine conversations in an open-domain information retrieval system. The structure of the dialog is simple because the project targeted the questioning-answering between the user and the system. Therefore, this corpus is easy to apply to the task-oriented dialog system, but it is not appropriated to construct the multimodal chat-talk dialog systems because the interactions are very restricted.

In this paper, we focus on a one-on-one Japanese dialog between humans to capture the various behavior in open-domain conversation. In addition, we collect the clear audio signals and video clips as much as possible to achieve the multimodal dialog system considering both verbal and non-verbal information.

III. Construction of Multimodal Chat-talk Corpus

A. Collection of Multimodal Dialog Data

The recording environment is shown in Figure 1. Two participants entered two individual sound-proof chambers to collect speech sounds without overlap. The interaction of the participants was captured by dynamic microphones (AT4055) placed near the participants' mouth and video cameras (GoPro HERO7) on the monitors. The captured video and sound were presented on the monitor of the peer and headphones in the other chamber in real time. With this setup, there was almost no delay in the video or sound, and the interlocutors could converse naturally.

The purpose of the dialog was "building a relationship," and each participant talked with their partner about five specific topics in Japanese to become more friendly. We consider chatting about one's preferences and tastes would be appropriate to topic for the chat-talk with the system, and so we prepared 10 topics based on the "Work (or studies)" and "Tastes and interests" categories of "The self-disclosure questionnaire" [11]. Table II shows the prepared topics. When recording a dialog, the operator selected five topics from the prepared topics according to the participant's preference.

Nineteen Japanese university students (15 males and 4 females) participated in the dialog collection. One "session" was a dialog about one topic. Consequently, 95 sessions (about 10 hours) were collected.

B. Closeness in Dyads

Before recording, interlocutors were asked to answer three questions: 1) whether the participants are acquainted with their partner or not, 2) how long the participants have known their partner, and 3) how much the participants feel a closeness to their partner (i.e., subjective closeness). For question 3), the participants rated the score on a 5-grade scale, from one (not at all) to five (very much). The questions 2) and 3) were asked to only acquainted pairs. Here, the mean and standard error of question 2) were 0.877±0.316 [year]. The mean and standard error of the scores of question 3) were 4.00±0.161. This reflects that many of the pairs in the acquaintance group had a close relationship.

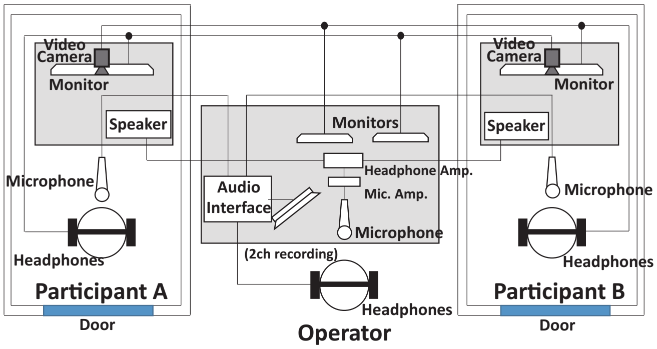

Fig. 1: Recording environment

C. Transcription and Annotation of Dialog Act

Five crowd-workers transcribed the collected dialog data, and the first author revised orthographical variants and punctuation mistakes, and we finally obtained 19,303 utterances. Then, we annotated a dialog act tag to each utterance. We prepared 12 dialog act tags based on tag sets of SWBD-DAMSL [12], JAIST annotated corpus [13], and listening-oriented dialog corpus [14]. The selected 12 tags are shown in Table III. The currently tag set only contains a minimum set to describe the collected dialog. We are going to extend the tag set to capture a more detailed dialog act sequence in the future study.

IV. Analysis of Dialog Act based on the Closeness in Dyads

In this study, we investigate how the tendency of the dialog acts occurrence differs under two closeness levels (stranger and acquaintance) to obtain the dialog model at the intention level considering the speaker closeness. Table IV summarizes the dialog data for the analysis. The analysis was conducted session by session.

A. Relative Frequency of Dialog Acts

First, we compared the frequency of the dialog acts between different closeness levels. Table V shows the mean and the standard error of the frequency of each dialog act. The table also shows the results of the unpaired t-test between two closeness levels. Significant differences were observed for self-disclosure, question, and request. The result shows that the speakers of the acquaintance groups did not talk about themselves comparing with that of the stranger group. This is consistent with social penetration theory, in which depth of self-disclosure changes from shallow to deep as the relationship develops. Many of the prepared topics in this study encouraged shallow self-disclosure by the participants (i.e., their preferences or tastes). Conversely, the speakers of the acquaintance group made request and disagree utterances more frequently. It is considered that the speaker of the acquaintance group did not hesitate to make such requests. This result implies that the speaker tended to make utterances that reflected his intention more directly after building the relationship. On the other hand, the speakers of the stranger group asked questions more frequently. This result suggests that the speaker tried to know each other in the initial stage of relationship.

B. Transition Probability of Dialog Acts

We investigated the transition of dialog acts. The transition probability of the dialog acts is calculated as follows:

$$P(a_{i+1}|a_i) = \frac{C(a_i, a_{i+1})}{C(a_i)} \quad (1)$$

where, a_i is the dialog act tag of the i-th utterance. $C(a_i)$ and $C(a_i, a_{i+1})$ are the frequencies of the dialog act a_i and

TABLE I: Comparison of characteristics between our target corpus and typical corpora

Name	Dialog participants	Dialog style	Diversity of the interaction	# participants	# words	Total duration	Closeness label
SBCSAE	Human	Multi-party	High	213	249k	23hours	Not included
Ritel Corpus	Human and machine	One-on-one	Restricted	13	60k	6hours	Not included
Our target corpus	Human	One-on-one	High	-	-	-	Included

TABLE II: Selected topics

Index	Topic
1	My favorite foods and beverages, and the ones I don’t like.
2	My favorite music, and the ones I don’t like.
3	My favorite reading matter.
4	My favorite movies and animations.
5	The best and the worst places that I have ever been to.
6	My tastes in clothing.
7	My favorite ways of spending spare time.
8	What I enjoy most, and get the most satisfaction from in my present school.
9	How I feel about my friends.
10	My strong and weak points for my work.

TABLE III: Dialog acts tag set and examples of utterances

Tag (abbreviation)	Description	Example
Self-disclosure (SD)	Utterance of disclosing one’s preference and emotion	I don’t like beans generally.
Information (INFO)	Utterance of conveying objective information	Well, that is the club for climbing.
Commit (CMT)	Utterance of suggesting something to partner	You should watch that.
Request (RQ)	Utterance of requesting something to partner	Stop it!
Confirmation (CFM)	Utterance of confirming	Is that so?
Question (QUES)	Utterance that expects a response from partner	Did you watch that movie?
Sympathy (SYM)	Utterance of agreeing and sympathizing to partner’s opinion	Indeed.
Disagree (DIS)	Utterance of disagreeing to partner’s opinion	I think that doesn’t suit me.
Backchannel (BC)	Utterance of encouraging partners utterance	yeah, yeah.
Admiration (AM)	Utterance of expressing surprising and admiration	Really?
Filler (F)	Filler words	Well...
Others (OTR)	Utterance that is not included in above categories	A sea squirt is (stop speech)

TABLE IV: Summary of analytical dialog data

Closeness	# pairs	# sessions	# utterances
Stranger	8	40	7218
Acquaintance	11	55	12085

TABLE V: Frequency of dialog acts in the stranger and acquaintance groups [%] (mean ± standard error).

Dialog act	Stranger	Acquaintance	p-value
Self-disclosure	21.07±0.96	17.88±0.67	0.008**
Information	19.42±1.11	22.13±0.99	0.072†
Commit	0.27±0.08	0.38±0.08	0.315
Request	0.06±0.03	0.19±0.05	0.024*
Confirmation	7.28±0.52	7.47±0.42	0.776
Question	8.78±0.54	6.88±0.44	0.008**
Sympathy	11.53±0.78	13.27±0.59	0.079†
Disagree	0.45±0.10	1.04±0.28	0.050†
Backchannel	17.22±1.04	17.93±1.24	0.663
Admiration	6.55±0.56	5.33±0.37	0.073†
Filler	6.20±0.60	5.93±0.37	0.70
Others	1.16±0.16	1.54±0.20	0.131

$^{\dagger}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01$

the dialog act transition from a_i to a_{i+1} in the session, respectively.

Figures 2 and 3 show the transition probability matrix of each group. The matrices show the average transition probability of sessions. In the figures, the transitions between the same person and different persons are treated individually. The figures show the trend of the dialog act transition is similar between groups.

TABLE VI: Frequency of 10 most-frequent pairs of dialog acts (mean ± standard error). The order of the items was decided by the mean of the average frequency of two groups. All items are transitions between different persons. Name of the dialog act tag is denoted in abbreviation. (see Table III).

Current→Next	Stranger	Acquaintance	p-value
RQ→SD	0.500±0.289	0.183±0.093	0.361
QUES→SD	0.391±0.031	0.289±0.022	0.009**
CFM→SYM	0.322±0.031	0.313±0.022	0.809
BC→SD	0.332±0.027	0.296±0.017	0.265
BC→ST	0.263±0.025	0.287±0.019	0.652
OTR→ST	0.233±0.061	0.203±0.039	0.674
QUES→ST	0.223±0.030	0.208±0.017	0.652
ADM→SD	0.235±0.032	0.163±0.014	0.046*
CMT→SYM	0.236±0.113	0.145±0.072	0.505
SD→BC	0.192±0.014	0.180±0.014	0.543

$^{\dagger}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01$

Here, we compared frequently occurred dialog act pairs. Table VI shows the top 10 most frequent dialog act pairs with respect to the mean of the average frequency of two groups. All items in the table are the transition from one speaker to another speaker. In particular, there was a significant difference at the 1% level in terms of “transition from question to self-disclosure.” This indicates that the typical questioning-answering frequently occurs in a talk between strangers, but decreases after the relationship has developed.

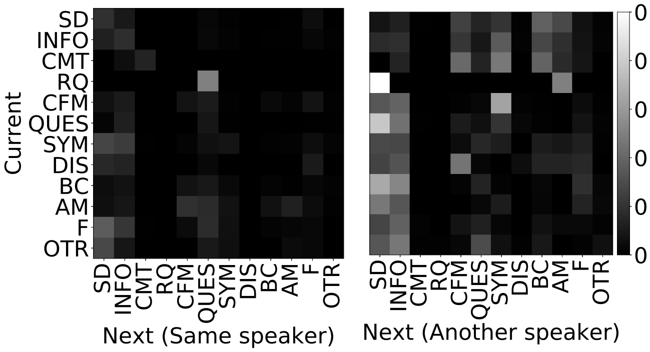

Fig. 2: Transition probability matrix of dialog acts in the stranger group.

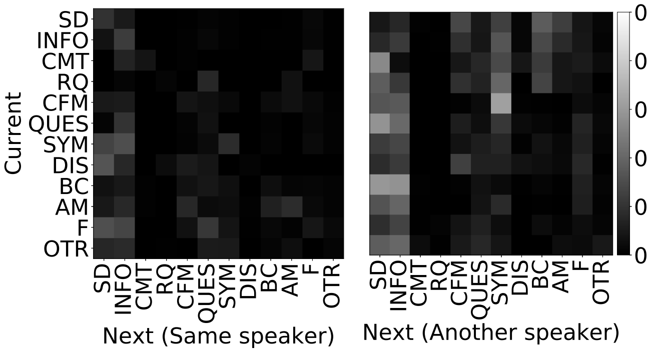

Fig. 3: Transition probability matrix of dialog acts in the acquaintance group.

C. Application to Dialog System

From the results, it is suggested that the collected corpus is useful to obtain the dialog model considering the speaker closeness because the tendency of the dialog acts occurrence partly differs depending on the closeness level. The analysis shows that the target dialog system should conduct simple questioning-answering in the initial stage of a relationship. On the other hand, the system needs to conduct the dialog under the assumption that the user continues speaking after the question when the relationship is developed.

V. Conclusion

In this paper, we constructed the multimodal Japanese chat-talk corpus considering the closeness in dyads. The results of the analysis suggested that the interlocutors frequently conducted simple questioning-answering with self-disclosure in the initial stage of the relationship. On the other hand, the speakers tended to make the request and disagree utterances in the conversation with the close friend. In future studies, we will collect the dialog of 100 participants to make a large scale multimodal chat-talk corpus. The dialog tag set will be also extended to capture the dialog act sequence in more detail. In addition, we will apply the knowledge obtained from the analysis in this paper to the dialog systems and conduct the user evaluation.

Acknowledgment

This work was supported by JST COI Grant Number JP-MJCE1303 and JSPS KAKENHI Grant Number JP18K18136.

References

[1] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[2] Alan Ritter, Colin Cherry, and Bill Dolan, "Unsupervised modeling of twitter conversations," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2010, HLT '10, pp. 172–180, Association for Computational Linguistics.

[3] Dan Bohus and Eric Horvitz, "Managing human-robot engagement with forecasts and... um... hesitations," in *Proc. the 16th International Conference on Multimodal Interaction*. 2014, ICMI '14, pp. 2–9, ACM.

[4] Pierrick Milhorat, Divesh Lala, Koji Inoue, Tianyu Zhao, Masanari Ishida, Katsuya Takanashi, Shizuka Nakamura, and Tatsuya Kawahara, "A conversational dialogue manager for the humanoid robot ERICA," in *Advanced Social Interaction with Agents*, pp. 119–131. Springer, 2019.

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[6] Yukiko Kageyama, Yuya Chiba, Takashi Nose, and Akinori Ito, "Improving user impression in spoken dialog system with gradual speech form control," in *Proc. SIGDIAL*, 2018, pp. 235–240.

[7] Yunkyung Kim, Sonya S. Kwak, and Myungsuk Kim, "Am I acceptable to you? Effect of a robot's verbal language forms on people's social distance from robots," *Computers in Human Behavior*, vol. 29, pp. 1091–1101, 2012.

[8] Irwin Altman and Dalmas A Taylor, *Social penetration: The development of interpersonal relationships*, New York: Holt, Rinehart & Winston, 1973.

[9] John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey, "Santa Barbara corpus of spoken American English, Parts 1-4," Linguistic Data Consortium.

[10] Sophie Rosset and Sandra Petel, "The Ritel Corpus - An annotated Human-Machine open-domain question answering spoken dialog corpus," in *Proc. the Fifth International Conference on Language Resources and Evaluation LREC*, Genoa, Italy, May 2006, European Language Resources Association (ELRA).

[11] Sidney M Jourard and Paul Lasakow, "Some factors in self-disclosure," *The Journal of Abnormal and Social Psychology*, vol. 56, no. 1, pp. 91–98, 1958.

[12] Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual," Tech. Rep., Institute of Cognitive Science, 1997.

[13] Kiyoaki Shirai and Tomotaka Fukuoka, "JAIST annotated corpus of free conversation," in *Proc. LREC*, 2018, pp. 741–748.

[14] Toyomi Meguro, Ryuichiro Higashinaka, Kohji Dohsaka, Yasuhiro Minami, and Hideki Isozaki, "Analysis of listening-oriented dialogue for building listening agents," in *Proc. SIGDIAL*, 2009, pp. 124–127.