Collection of Multimodal Dialog Data and Analysis of the Result of Annotation of Users' Interest Level

Masahiro Araki[1], Sayaka Tomimasu[1], Mikio Nakano[2], Kazunori Komatani[3], Shogo Okada[4], Shinya Fujie[5], Hiroaki Sugiyama[6]

[1]Kyoto Institute of Technology, [2]Honda Research Institute Japan Co., Ltd., [3]Osaka University,
[4]Japan Advanced Institute of Science and Technology [5]Chiba Institute of Technology,
[6]Nippon Telegraph and Telephone Corporation
araki@kit.ac.jp, tomi@ii.is.kit.ac.jp, nakano@jp.honda-ri.com, komatani@sanken.osaka-u.ac.jp,
okada-s@jaist.ac.jp, shinya.fujie@p.chibakoudai.jp, sugiyama.hiroaki@lab.ntt.co.jp

Abstract

The Human-System Multimodal Dialogue Sharing Corpus Building Group is acting as a working group of SIG-SLUD for the purpose of constructing a corpus for evaluating elemental technologies of the multimodal dialogue system. In this paper, we report the results of recording chat dialogue data between a human and a virtual agent by the Wizard of OZ method conducted in 2016, and the results of the analysis of annotations of users' interest level in the data.

Keywords: Multimodal dialogue corpus, users' interest level, dialogue system

1. Introduction

Recently, multimodal dialogue systems that utilize not only spoken language but also image and other modalities are getting much attention, thanks to improvements in language, speech, and image processing techniques and their underlying machine learning technologies as well as the advancement of hardware such as computers, sensors, display, and robots. However, how such systems should use multimodal information when communicating with humans is still under investigation. One of the reasons is that there are not enough shared multimodal dialogue data with annotations referring to users' intentions, emotions, attitudes, and dialogue situations. Since a lot of effort is needed for annotation, it is desirable that multiple research institutions collaborate to annotate.

With this background, we initiated an activity to build a shared corpus of human-system multimodal dialogues[1]. So far, a number of shared corpora of multimodal dialogues among humans have been built and there has been plenty of work on the analysis of these corpora (e.g., (Carletta, 2007; Janin et al., 2003)). In contrast, the goal of this project is to contribute to component technology development for multimodal dialogue systems. Therefore, we focused on analyzing how humans behave toward multimodal dialogue systems, rather than analyzing human-human dialogues. In human-system dialogues, user behaviors are different from human behaviors in human-human dialogues because the users realize that they are talking to a system. By collecting data of human-system dialogues, we can analyze user behaviors to contribute to improve components of dialogue systems.

We collected multimodal dialogues between a user and a virtual agent, which was operated by the Wizard of Oz method, and annotated them with labels indicating whether the user is interested in the topic or not. We used a virtual agent because we consider virtual agents and robots are promising natural user interfaces as users would feel easier in talking to them than to devices without human-like characters, and virtual agents are easier to use in data collection than robots.

This paper reports our multimodal dialogue data collection protocol (Sec. 2.) and annotation method (Sec. 3.), and discusses future data collection and annotation through the results of the analysis of annotations (Sec. 4.). Then, by considering the differences with related work (Sec. 5.), we list issues related to sharing the corpus that we will build with the research community (Sec. 6.).

2. Multimodal dialogue corpus

We implemented an environment for recording multimodal dialogue between a human and a virtual agent (Tomimasu and Araki, 2016). The virtual agent [2] was manipulated by an operator via the Wizard of Oz (WoZ) method, which facilitates data collection. A human so-called "Wizard" simulates a dialogue system that interacts with the human users via the GUI interface shown in Figure 1. The interface features a topic selector, a selectable list of typical utterances for each topic, and a selectable list of general responses in chat dialogue.

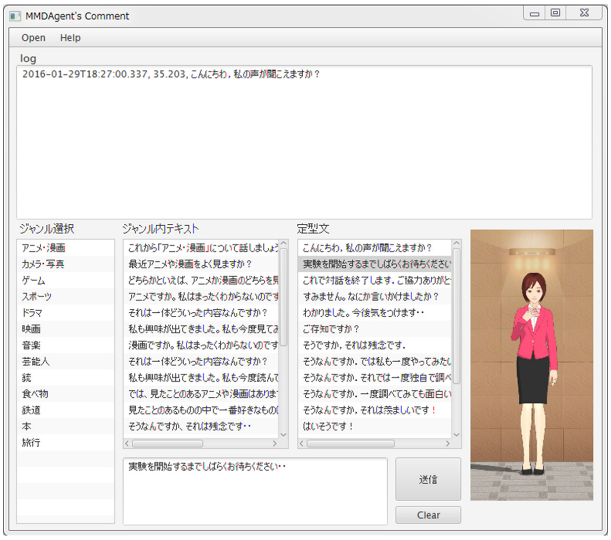
Before the data collection, the participants assessed their interest in 12 topics. During the data collection, six topics were selected as dialogue theme from both the favorite and non-favorite groups of topics. The operator tried to follow the dialogue engagement of the participants by selecting the utterance with the different initiative. The operator also limited the number of exchanges for each topic to approximately 10. The behavior of the participants was recorded via a video camera and Microsoft Kinect ®V2 sensor.

An example of the dialogue is shown in Figure 2 (it is originally Japanese that translated to English.).

In April 2016, data from four people (hereafter referred to as data 1) were experimentally recorded and the problems in the recording environment were examined. At this time,

[1] This activity is being conducted by the Human-System Multimodal Dialogue Sharing Corpus Building Group of SIG-SLUD of the Japanese Society for Artificial Intelligence.

[2] http://www.mmdagent.jp/



2016-01-29T18:27:00.337, 35.203, こんにちわ，私の声が聞こえますか？

アニメ・漫画
カメラ・写真
ゲーム
スポーツ
ドラマ
映画
音楽
芸能人
銃
食べ物
鉄道
本
旅行

これから「アニメ・漫画」について話しましょう
最近アニメや漫画をよく見ますか？
どちらかといえば、アニメか漫画のどちらを見
アニメですか。私はまったくわからないのです
それは一体どういった内容なんですか？
私も興味が出てきました。私も今度見てみ
漫画ですか。私はまったくわからないのです
それは一体どういった内容なんですか？
私も興味が出てきました。私も今度読んで
では、見たことのあるアニメや漫画はありま
見たことのあるものの中で一番好きなもの
そうなんですか、それは残念です・・

こんにちわ，私の声が聞こえますか？
実験を開始するまでしばらくお待ちください
これで対話を終了します．ご協力ありがと
すみません。なにか言いかけましたか？
わかりました。今後気をつけます・・
ご存知ですか？
そうですか，それは残念です．
そうなんですか．では私も一度やってみたい
そうなんですか．それでは一度独自で調べ
そうなんですか．一度調べてみても面白い
そうなんですか，それは羨ましいです！
はいそうです！

実験を開始するまでしばらくお待ちください・・

Copyright 2009-2013 Nagoya Institute of Technology (MMDAgent Model "Mei")

Figure 1: GUI for Wizard.

```
S: Let's talk about railway.
U: Yes.
S: Do you like trains?
U: Well, not so much.
S: Then do you often take a train?
U: I have not taken the train recently.
S: In what kinds of situations do you ride
   the train?
U: Well, when going out of the prefecture
   ... and when I go out for a drink.
   I cannot use my motorcycle.
```

Figure 2: An example of dialogue

a trial annotation of participants' interest level in the topics was carried out by members of the working group. We made improvements such as randomizing the order of presenting topics and clarifying instructions to participants; in January 2017 data from 10 people (hereafter data 2) were recorded.

Regarding data 2, after the preliminary annotation by three annotators and improvement of the annotation manual based on their work, a release version of the annotation was created by another three annotators.

In the annotation of chat dialogue between a person and a dialogue system, we assigned one of the following labels to each exchange: interest (o), unknown (t), and no interest (x) to each exchange (i.e., a pair of "system utterance S" and "user utterance U"). The assessment was done by considering the various sources of participant's information, such as facial expressions, prosody, speech content, etc.

3. Annotation of users' interest level

The estimation of users' interest level in a dialogue topic enables the system to capture the users' preferences, which is essential when developing user-adapted dialogue systems. In this section, we analyze our corpus with the annotation of users' interest level to examine the reliability of the annotations and to improve the annotation manual.

Table 1: Annotation results

Label	First	Second
Interested	907	992
Unknown	162	267
Not interested	1276	1108
Error	22	0
Fleiss' kappa	0.407	0.490

The specific procedure of the analysis is as follows. At first, three graduate students studying human-computer interaction annotated data 2 (10 persons, 789 interactions) with very intuitive instructions. Through the analysis and discussion about differently judged interactions, we developed an annotation manual. Then we compared the first round of annotations with the annotations carried out by three new annotators, who received the annotation manual for instruction.

In the first round of annotations, the graduate students intuitively judged whether the participant was interested in the topic or not. It should be noted that they judged the presence of the interest in the current dialogue topics, not whether the participants enjoyed the dialogue or not. The column labeled *First* in Table 1 shows the results of the first annotation. Here, the simple hearing back actions were annotated as errors.

Since we controlled the number of *interested* and *not interested* topics based on the preliminary survey, we expected the number of annotated labels (*interested* and *not interested*) to be almost the same; however, the number of *not interested* labels is higher. In the first annotation, Fleiss' kappa was 0.407, which is interpreted as between Fair agreement and Moderate agreement.

We analyzed the criteria of the annotators in cases where they did not agree and looked for possible causes of the disagreement as follows.

- At the beginning of the dialogue, some annotators tried to judge a dialogue as *interested* or *not interested* using little information, but others thought those interactions should be judged as *unknown*.

- Some annotators judged interactions based on only part of the interactions, others based their judgment on the entire interaction.

- Some annotators always utilized single or a few modalities such as smile instead of the contents of the interaction.

- Some annotators consider that the criteria of the labels are consistent among participants, but the others consider that the criteria differ for each participant.

According to the results of this analysis, we developed an annotation manual with the indications that follow, and had three annotators who work at another research institute carry out a second round of annotations.

- The beginning of the dialogue should be labeled as *unknown* if you are not completely sure.

- The judgment should be done based on the whole interaction instead of a part of it.
- Annotators firstly watch the whole video and understands the participant's habits of emotional expressions before the judgment.
- The interaction should be labeled *Error* when the interaction is a completely unexpected one.

The column labeled *Second* in Table 1 shows the results of the second round of annotations.

In the second annotation, Fleiss' kappa improved to 0.490, which is interpreted as Moderate agreement.

Our experiment shows that, even with a difficult problem such as judging the presence or absence of interest in an unfamiliar situation as in human-system dialogue, we can obtain reliable annotations with a well-designed annotation manual.

4. Analysis of label distribution

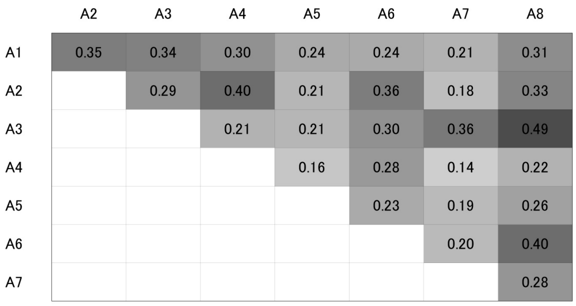
The subjectivity of coders influences the annotation of participants' interest level in the dialog. Data 1 was collected for analysis of the subjectivity of coders and the label distribution. In this section, we analyze the distribution of annotated labels by calculating the level of agreement (κ coefficient) between coders of data 1. A total of eight coders annotated the interest level for four participants who participated in the experiments. For each exchange (interaction), the coders were asked to annotate the labels: interest (o), unknown (t), and no interest (x).

Fleiss's κ (κ_f) was calculated as the agreement between the eight coders. The agreement was 0.26, which is considered low. The agreement between each pair of coders was calculated to analyze the reasons for the low agreement. Cohen's κ (κ_c) was calculated for each pair of eight coders (A1-A8). Figure 3 shows the matrix of κ_c among coders. Although the κ_c was less than 0.3 in almost all cases, the agreement between some pairs was more than 0.4 (Moderate agreement). Furthermore, we analyzed the similarity of annotations between coders through hierarchical clustering with Ward's method using κ_c as a distance measure. The similarity between one coder and the other coders was also calculated as the average of κ_c based on the matrix (Figure 3).

Figure 4 shows the dendrogram that denotes the clustering results and average κ_c. From the figure, coders A5 and A7 annotated interest level in a different manner compared to the other coders because the averages of A5 and A7 were the lowest and second lowest, respectively, among all the coders. The clustering analysis results show that the annotation task was influenced by the coders' subjectivity. It also clarified the dissimilarity of label distributions among coders.

5. Related work

For analyzing human-human multimodal conversations, several meeting conversation corpora with multiple participants have been released and shared such as the AMI (Augmented Multi-party Interaction) corpus (Carletta, 2007) and the ICSI meeting corpus (Janin et al., 2003). The CHIL



	A2	A3	A4	A5	A6	A7	A8
A1	0.35	0.34	0.30	0.24	0.24	0.21	0.31
A2		0.29	0.40	0.21	0.36	0.18	0.33
A3			0.21	0.21	0.30	0.36	0.49
A4				0.16	0.28	0.14	0.22
A5					0.23	0.19	0.26
A6						0.20	0.40
A7							0.28

Figure 3: Cohen's kappa (κ_c) between each pair of coders

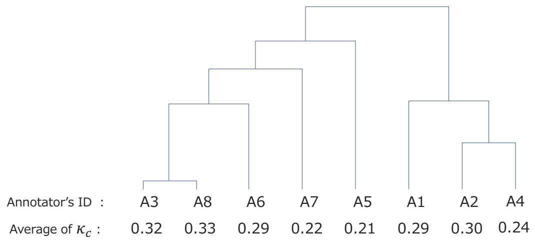


Figure 4: Clustering results of annotators based on Cohen's kappa (κ_c)

(Computers in Human Interaction Loop) treats human-human interactions in offices and classrooms (Waibel and Stiefelhagen, 2009), and VACE (Video Analysis and Content Extraction) treats human-human interactions in battle-game sessions in the air force (Chen et al., 2006). A corpus of political debates in a TV program has also been shared for analyzing social interactions (Vinciarelli et al., 2009).

Several multimodal corpora, which are not those of interactions, have also been published, such as the ones for assessing public speaking ability and anxiety (Chollet et al., 2016) and for recognizing group-level emotion on in-the-wild data (Dhall et al., 2017). Systems based on such multimodal analyses have been constructed; for example, a system that analyzes nonverbal human behaviors was developed and used to assess indicators of psychological distress, such as depression and post-traumatic stress disorder (PTSD) (Stratou and Morency, 2017).

Our goal in this project is to contribute to component technology development for multimodal dialogue systems. Therefore, our target is to construct dialogue corpora; more specifically, those are not of human-human dialogues but of human-system dialogues. The most important difference between human-human and human-system dialogues is whether users realize that they are talking to a system. User behaviors differ when they talk to a system or a human. Data on human-system dialogues contain real behaviors of users, which can be used to predict user actions in actual dialogue systems. Furthermore, dialogues have inherent characteristics such as the fact that verbal interactions are composed of multiple turns and the existence of certain dialogue states. In the WoZ system currently used for our data collection, dialogues can be divided by topics presented by the system, which can be regarded as one of the dialog states. Collecting such dialogue data can lead to a novel system design that will not easily bore users by considering such states.

In terms of sharing the human-system dialogue corpus, there is an ongoing project to collect and share a text chat corpus, which also conducts shared tasks using it (Higashinaka et al., 2015). Currently, a competition for speech-input chatbots, the Amazon Alexa Challenge[3], is also being conducted. The target of our project is not text-input or speech-input chatbots but multimodal dialogue systems.

Constructing dialogue robots is one of the ultimate goals of dialogue system research, and thus has been investigated by many researchers (Bohus and Horvitz, 2009; Al Moubayed et al., 2012; Sugiyama et al., 2015; Matsuyama et al., 2015; Lala et al., 2016). Some studies focused on the detection of user interests to enable a system to adapt topics to user preferences (Hirayama et al., 2011; Chiba et al., 2014; Tomimasu and Araki, 2016). Our corpus can be used for developing and improving such systems.

6. Issues on privacy and future work

The collected multimodal dialogue data includes the personal information of participants, such as their faces and voices. For the purpose of privacy protection, several fundamental rules for treating personal data have been established by the Japanese government. Thus, the following issues should be addressed prior to sharing the corpus, even for research purposes.

- Careful discussions are required on what kind of research will be done prior to data collection since the participants have to sign an agreement on the range of the data use.
- It is preferable that participants be allowed to reject the agreement even after they have signed it. Thus, the data should be distributed by an agency that specializes in research data distribution.
- Since crowdsourcing cannot be used for data annotation because of these limitations, the research group should be maintained continuously to preserve the quality of the collected data.

Currently, the working group is discussing the abovementioned issues in order to release the collected data. We are planning to share our corpus with the annotation results by multiple annotators, in the same way as the text chat corpus was shared (Higashinaka et al., 2015). Since the user's interest level to be annotated in our project is based on the subjective judgments of annotators, the annotation process is not to give a reference label that can be uniquely determined, and thus differences among individual annotators are inevitable. Therefore, the corpus will also be used for studying how to handle such subjective annotation results and how to apply such labels to dialogue system research.

[3] https://developer.amazon.com/alexaprize

7. Bibliographical References

Al Moubayed, S., Skantze, G., Beskow, J., Stefanov, K., and Gustafson, J. (2012). Multimodal multiparty social interaction with the furhat head. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI)*, pages 293–294.

Bohus, D. and Horvitz, E. (2009). Models for multiparty engagement in open-world dialog. In *Proceedings of Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 225–234.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T. X., Tu, J., Huang, Z., Harper, M., Quek, F., Xiong, Y., McNeill, D., Tuttle, R., and Huang, T. (2006). Vace multimodal meeting corpus. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction (MLMI05)*, pages 40–51.

Chiba, Y., Ito, M., Nose, T., and Ito, A. (2014). User modeling by using bag-of-behaviors for building a dialog system sensitive to the interlocutor's internal state. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 74–78, June.

Chollet, M., Wortwein, T., Morency, L.-P., and Scherer, S. (2016). A multimodal corpus for the assessment of public speaking ability and anxiety. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., and Gedeon, T. (2017). From individual to group-level emotion recognition: EmotiW 5.0. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages 524–528, New York, NY, USA. ACM.

Higashinaka, R., Funakoshi, K., Araki, M., Tsukahara, H., Kobayashi, Y., and Mizukami, M. (2015). Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 87–95, September.

Hirayama, T., Sumi, Y., Kawahara, T., and Matsuyama, T. (2011). Info-concierge: Proactive multi-modal interaction through mind probing. In *The Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2011)*.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings of IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, pages I–364–I–367, April.

Lala, D., Milhorat, P., Inoue, K., Zhao, T., and Kawahara, T. (2016). Multimodal interaction with the autonomous android ERICA. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, pages 417–418.

Matsuyama, Y., Akiba, I., Fujie, S., and Kobayashi, T. (2015). Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech & Language*, 33(1):1 – 24.

Stratou, G. and Morency, L.-P. (2017). Multisense—context-aware nonverbal behavior analysis framework: A psychological distress use case. *IEEE Transactions on Affective Computing*, 8(2):190–203, April.

Sugiyama, T., Funakoshi, K., Nakano, M., and Komatani, K. (2015). Estimating response obligation in multi-party human-robot dialogues. In *Proceedings of IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 166–172, Seoul, South Korea.

Tomimasu, S. and Araki, M. (2016). Assessment of users' interests in multimodal dialog based on exchange unit. In *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, MA3HMI '16, pages 33–37, New York, NY, USA. ACM.

Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009). Canal9: A database of political debates for analysis of social interactions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–4, Sept.

Waibel, A. and Stiefelhagen, R. (2009). *Computers in the Human Interaction Loop*. Springer Publishing Company, Incorporated, 1st edition.