Small Talk Improves User Impressions of Interview Dialogue Systems

Takahiro Kobori[†], **Mikio Nakano**[‡], **and Tomoaki Nakamura**[†]
[†]University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan
{t_kobori@radish, naka_t@apple}.ee.uec.ac.jp
[‡]Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako, Saitama 351-0188, Japan
nakano@jp.honda-ri.com

Abstract

This paper addresses the problem of how to build interview systems that users are willing to use. Existing interview dialogue systems are mainly focused on obtaining information from users, thus they just repeatedly ask questions. We propose a method for improving user impressions by engaging in small talk during interviews. The system performs frame-based dialogue management for interviewing and generates small talk utterances after the user answers the system's questions. Experimental results using a text-based interview dialogue system for diet recording showed the proposed method gives a better impression to users than interview dialogues without small talk. It is also found that generating too many small talk utterances makes user impressions worse because of the system's low capability of continuously generating appropriate small talk utterances.

1 Introduction

Our goal is to build dialogue systems that can obtain information from users. In this paper, we call such systems *interview dialogue systems*. An example is a dialogue system that interviews a user about what he/she ate and drank. The information obtained by the system is expected to be used for health care.

Although interviews have not been as popular as database search and reservations as applications of dialogue systems, they have commercial potential (Stent et al., 2006). Interview dialogue systems would be useful not only because they save human labor but also because users are expected to disclose their personal information to automated systems more often than to human-operated systems (Lucas et al., 2014).

We propose a method for dialogue management for such a dialogue system. Although several interview dialogue systems have been developed so far, most of them put their focus mainly on obtaining information, repeating questions and making mechanical dialogues. It might be acceptable if the user is expected to use the system only once, like a system for an opinion poll. However, such a strategy is not acceptable for systems like the one for diet recording, because users might not want to use such a system every day.

In human-human conversations, participants sometimes try to obtain information from another participant while enjoying the chat. If a system can engage in such kinds of conversation, a user may be willing to use it. However, the capability of even state-of-the-art chat systems is not good enough to chat for a long time. They sometimes cause dialogue breakdowns for various reasons (Higashinaka et al., 2015).

Our proposed dialogue management method mainly engages in an interview dialogue and sometimes inserts *small talk utterances*.[1] In this paper, a small talk utterance means an utterance that is not directly related to the task of the dialogue but makes the dialogue smoother and friendly. Examples of small talk utterances are utterances telling impression (e.g., "It sounds very nice") and self-disclosures (e.g., "That's my favorite food."). We expect that generating small talk utterances will enable users to enjoy using the system and they will want to use the system again.

Using the proposed method, we built an interview dialogue system for diet recording and con-

[1]We use the term *utterance* rather than *sentence* even though we deal with only text-based dialogue systems in this paper, because sentences used in those systems are more colloquial.

Proceedings of the SIGDIAL 2016 Conference, pages 370–380,
Los Angeles, USA, 13-15 September 2016.

ducted a user study to investigate the effectiveness of the small talk utterances. We found that the small talk utterances give the user a better impression but it was suggested that generating too many small talk utterances increases the possibility of generating unnatural utterances, resulting in bad impressions.

This paper is organized as follows. Section 2 surveys related work, and Section 3 proposes the method for dialogue management. Section 4 explains in detail the interview dialogue systems for diet recording as an implementation of the proposed method. Section 5 shows the experimental evaluation results before concluding the paper in Section 6.

2 Related Work

Although interviews have not been popular applications of dialogue systems, several systems have been developed so far.

One of the earliest systems is MORE (Kahn et al., 1985), which can elicit knowledge for diagnosis from human experts. It uses a number of heuristic rules to generate questions to human experts. Although the paper does not clearly state how it understands user replies, it does not seem to perform complicated language processing. Stent et al. (2006) built a spoken dialogue system for interview-based surveys for rating college courses. They showed dialogue epiphenomena can be used to learn more than the system asks. Johnston et al. (2013) built a spoken dialogue system for government and social scientific surveys. They are concerned with confirmation strategies for reducing errors in the surveys. Skantze et al. (2012) use robot behaviors for increasing the reply rate in survey interviews. All these systems focus on obtaining information from users. They are suitable to be used once but it is not clear whether users want to continuously use them.

On the contrary, *chat-oriented dialogue systems*, which can engage in small talk, have been built so that users will enjoy conversations with them (Wallace, 2008; Wilks et al., 2011; Higashinaka et al., 2014). It has been tried to combine chat-oriented dialogue systems with task-oriented dialogue systems (Traum et al., 2005; Nakano et al., 2006; Lee et al., 2006). Recent commercial dialogue systems such as Siri (Bellegarda, 2013) also have functionality for engaging in small talk.

Incorporating small talk into interview dialogue systems has been considered as well, since small talk is known to be effective in building *rapport* (Bickmore and Picard, 2005), they are expected to increase the rate that the user honestly answers the questions. For example, Conrad et al. (2015) showed that small talk in survey interviewing to increase the users' comprehension and engagement. Bickmore and Cassell (2005) also used small talk to increase trust. Unlike those studies whose aim is to obtain more information from the users, we focus on how to give better impressions to the users. In addition, while both Conrad et al. (2015) and Bickmore and Cassell (2005) conducted Wizard-of-Oz based studies, we take into account that it is inevitable for systems to generate inappropriate utterances.

3 Proposed Method

There are two possible dialogue management strategies for engaging in both interview dialogues and chat-oriented dialogues. One is to deal with chat as the primary strategy and sometimes invoke an interview dialogue to ask questions to the users. This strategy is taken by some of the previously built dialogue systems that integrate task-oriented dialogues and chat-oriented dialogues (Nakano et al., 2006; Lee et al., 2006). The other strategy is to deal with interviewing as the primary strategy and chat as the secondary strategy.

In the former approach, since the capability of the current chat-oriented dialogue systems is not good enough to always generate utterances that match the dialogue context (Higashinaka et al., 2015), engaging in chat for many turns might make the user's impression worse.

We therefore take the latter approach. Our method systematically asks questions for the interview based on frame-based (Bobrow et al., 1977; Goddeau et al., 1996), agenda-based (Bohus and Rudnicky, 2009), or other kinds of dialogue management. Then when the user replies to the system's questions, it may start small talk by choosing one of the small talk utterances stored in a database. After several turns, it goes back to the interview. When to start small talk and when to finish are determined by heuristic rules or probabilistic rules learned from a corpus. By this strategy, even if small talk does not go well, the system can go back to the interview and evolve the dialogue.

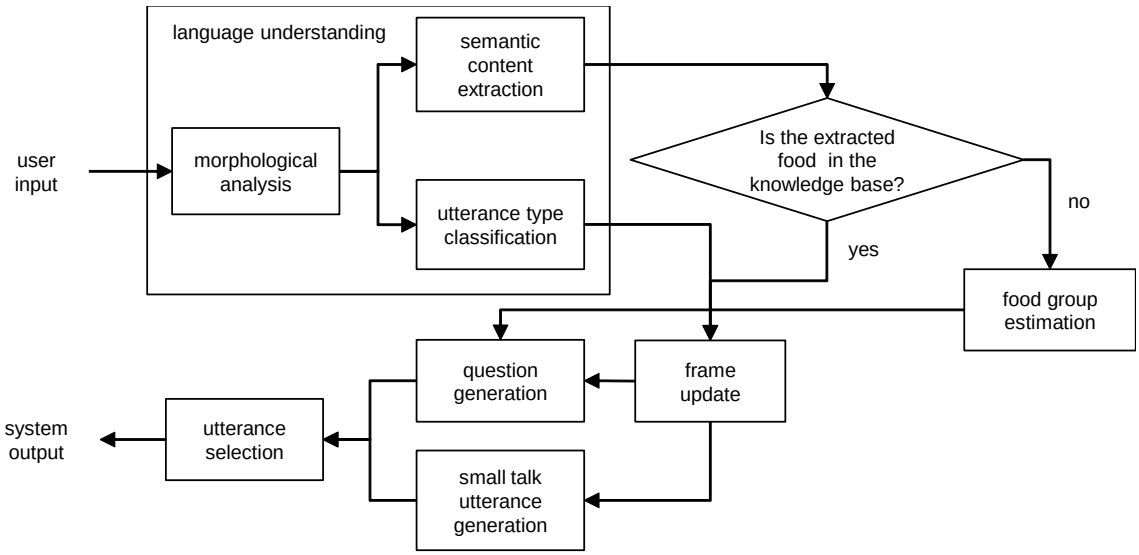

Figure 1: Architecture for the interview dialogue system for diet recording

4 Implementation: An Interview Dialogue System for Diet Recording

Based on the proposed method, we have developed a Japanese text-based interview dialogue system that asks the user what he/she ate and drank the day before. Figure 1 shows the architecture of the system.

Note that the goal of the system is to obtain rough information of what the user had each day. We assume the information is used to know the tendency of the user's dietary habits. Obtaining detailed dietary records so that it can be used for nutritional guidance is out of the scope of our research.

4.1 Knowledge Base

Our system assumes most users have meals with typical meal compositions for Japanese. For example, lunch can consist of a one-dish meal and soup, or it can consist of *shushoku* (side dish mainly containing carbohydrates), a couple of *okazu* (main or side dish containing few carbohydrates), and soup. Each kind of food can be one of these categories; for example, steamed rice and bread are *shushoku*, and sandwiches and tacos are one-dish meals. We call these categories *food groups*. The system has a knowledge base that contains a list of foods for each food group as shown in Table 1.

4.2 Understanding User Utterances

The language understanding module first performs a morphological analysis using MeCab (Kudo et al., 2004) to segment the input text into words and get their part-of-speech information.

It then determines the type of the user utterance. The type is either *greeting*, *affirmative utterance* (including replies to system questions), or *negative utterance*. The number of types is small because, in interview dialogues, user utterances have small variations. An utterance telling the food and drink the user had is an affirmative utterance. This utterance type classification is done by LR (Logistic Regression), which uses bag-of-words features. We used LIBLINEAR (Fan et al., 2008) for the implementation of LR.

It then performs semantic content extraction, that is, obtaining five kinds of information, namely, food and drink, ingredient, food group, amount of food, and time of having food. This is done by CRF (Conditional Random Fields) using the IOB2 tagging framework (Hahn et al., 2011). For the CRF, we used commonly used features such as unigram and bigram of the surface form, original form and part of speech of the word. We used CRFsuite (Okazaki, 2007) for the implementation of CRF.

These statistical models for LR and CRF were trained on 5,630 utterances. This set was artificially created by randomly changing content words in 563 sentences manually written by developers.

4.3 Dialogue Management for Interviewing

Dialogue management for interviewing is based on a frame. Slots of the frame are compositions

Food group	Examples instances	# of instances
shushoku (side dish mainly containing carbohydrates)	steamed rice, bread, cereal	20
okazu (main or side dish containing few carbohydrates)	Hamburg steak, fried shrimp, grilled fish	106
soup	corn soup, *miso* soup	18
one-dish meal	sandwich, noodle soup, pasta, rice bowl	78
drink	orange juice, coffee	32
dessert	cake, pancake, jelly	16
confectionery	chocolate, donut	34
total		304

Table 1: Content of the knowledge base

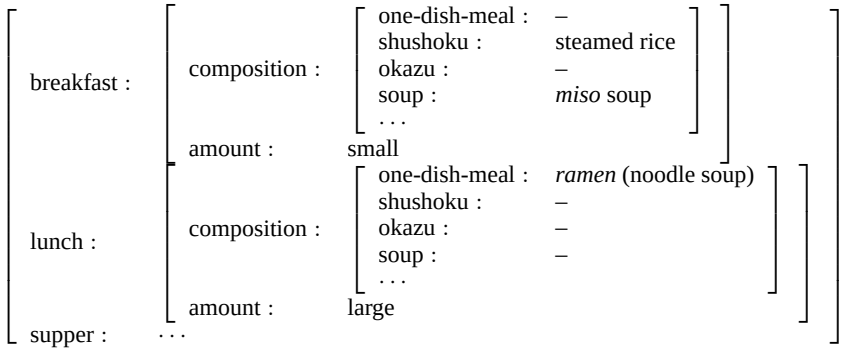

Figure 2: A snapshot of the frame

for each meal (breakfast, lunch, supper) and the amount of each food. Figure 2 shows a snapshot of the frame.

The frame is updated each time the user makes an utterance, based on its language understanding result. When there is food or drink in the understanding result, the system needs to know its group so that it can fill the appropriate slot of the frame. For example, when the user says he/she had steak for supper, the system needs to know if it is an *okazu* (main or side dish) so that it can fill the "okazu" slot of the "composition" slot of the "supper" slot. This is done using the food list in each food group in the knowledge base. If the food is not in the food and drink list, the system estimates its food group and requests confirmation from the user as will be explained in Section 4.4. Slot values can be a set of food and drink. So if the user says he/she had a steak and a salad, the *okazu* slot value is the set of "steak" and "salad".

The system-utterance selection is done with manually written rules. The system asks what the user ate and drank in order. This is because in human-human dialogues we collected in advance, participants asked what the other participant had in a particular order. In addition the system asks the user brief descriptions of the food, and then asks the composition in detail. For example, when asking about breakfast, the system asks first "what did you have for breakfast?" and then asks detailed questions such as "what else did you have?" and "what did you have for *shushoku*?" When the frame satisfies conditions for each meal (breakfast, lunch, and supper), the system moves to asking about the next meal, and then finishes after obtaining information about all meals. In this process, constraints on slot values are considered; for example, if the *one-dish-meal* slot value is not empty, the system does not ask about *shushoku*, because people do not tend to have both one-dish meals and *shushoku* in one meal. The system's questions are not always the same; they are randomly chosen from a variety of candidate expressions.

This frame representation is not perfect in that it cannot represent meal compositions that are not typical for Japanese users. Some users may have more than three meals in one day. Augmenting the system to deal with a variety of meal composition is among our future work.

Even if the system cannot understand the user's answer perfectly, the system moves the dialogue forward so that the dialogue does not get stuck.

4.4 Acquiring Food Groups

When the recognized food is not in the database, to estimate its group, we used a method proposed

S:	What did you have for breakfast?
U:	I had *natto-gohan* (steamed rice with fermented soybeans).
S:	Is *natto-gohan* an *okazu* or *shushoku*?
U:	It's a *shushoku*.

Figure 3: Example dialogue for food group acquisition

by Otsuka et al. (2013). Although they used both a model trained from a food database and Web search results, we only used the former. It estimates the group of the food as one of the seven groups in Table 1 and asks a question such as "Is *osuimono* (Japanese broth soup) soup?". This is done by logistic regression, which uses the bag of words, unigram and bigram of characters as features, the type of characters used in Japanese (*hiragana*, *katakana*, Chinese characters, and alphabet). The amount of training data consists of 863 expressions.

The system does not always ask back to the user only the top estimation result. It sometimes generates n-ary questions using n-best estimation results. For example, a binary question "Is sweet roll a confectionery or a one-dish meal?" can be asked. This is because the top estimation result is not always correct. In addition, n-ary questions are sometimes easy to understand because the user does not know the list of food groups in advance and he/she may not understand what *shushoku* really means. How many candidates are used in the question is decided based on posterior probabilities but we omit the detailed explanation because it is not really related to the main topic of this paper.

The dialogue management for acquiring the group of a food is performed separately from the management for interview dialogues; that is, when the food name that the user says is not in the database, the control moves to the food group acquisition dialogue managers, and after obtaining the food group, the control moves back to the interview dialogue manager. Figure 3 shows a translation of an example food group acquisition dialogue.

4.5 Generating Small Talk Utterances

Small talk utterances are selected from a predefined list based on the type and the content of the preceding user utterance. When the user utterance is affirmative, negative utterances are avoided as

Type	#
showing empathy	26
telling impression of that the amount is large	22
telling impression of that the amount is small	50
asking a question	6
self-disclosure	2
backchannel	6
telling impression of the user's negative answer	7
reaction to individual food	323
Total	442

Table 2: The number of small talk utterance candidates for each type

candidates for small talk utterances. For example, the following utterances can be chosen.

It's my favorite food! (self-disclosure)
Great! (showing empathy)
Is it tasty? (asking a question)

When the user utterance is negative, utterances showing impression of the user's negative answer, such as "It lowers my spirit...", "Let's go to eat something you like!" are selected.

In addition, when a specific food is included in the preceding user utterance, small talk utterances corresponding to the food are chosen. For example, if the user utterance includes "tiramisu", utterances such as "Tiramisu means 'cheer me up' in Italian." can be chosen.

When the amount of a meal is mentioned by the user, a small talk utterance showing an impression of the amount is chosen. For example, if the user says that the amount is large (e.g., "I ate too much"), an utterance such as "Please be careful about your health!" is chosen.

We constructed the predefined list of small talk utterances by asking our colleagues and 100 people recruited through a crowdsourcing site to list them, obtaining 442 utterances in total. The number of utterances for each type is shown in Table 2.

4.6 Response Selection

The response selection module needs to select one utterance from many candidates for questions and small talk utterances. Although learning a strategy for selecting appropriate system utterances from a corpus is among our future work, the current system adopts a very simple heuristic rule. First it decides if it asks questions or generates small talk utterances. If the number of small talk utterances generated after the previous system question is smaller than N, which is set in advance, it generates a small talk utterance, and otherwise it asks

$N = 0$ (no small talk utterances)

S:	What did you have for lunch?
U:	I had *okonomiyaki* (Japanese pancake) for lunch.
S:	Did you have any other food for lunch?
U:	No, I didn't have any other food.
S:	How much did you have for lunch yesterday?
U:	I had only a small amount.

$N = 1$ (one small talk utterance)

S:	What did you have for breakfast yesterday?
U:	*Shoku-pan* (sliced bread) and corn soup.
S:	Those are good!
U:	I agree.
S:	What else did you have for breakfast yesterday?

$N = 3$ (three small talk utterances)

S:	What did you have for lunch yesterday?
U:	I had a Hamburg steak.
S:	You had a Hamburg steak, did you?
U:	Yes, it was a Hamburg steak.
S:	Do you like Hamburg steaks?
U:	Yes, I like them very much.
S:	Actually I had a Hamburg steak.
U:	Really? Was it tasty?
S:	What else did you have for lunch yesterday?

Figure 4: Example dialogues
(S means the system and U means the user)

Favored condition	Frequency
NO	67
ONE	46
TWO	107
THREE	270
TWO-CONSECUTIVE	65
THREE-CONSECUTIVE	45
Total	600

Table 3: Result of questionnaire survey on the number of small talk utterances

a question. Small talk utterances are randomly selected from the candidates but repeating the same small talk utterance within the N turns is avoided.

4.7 Example Dialogues

Figure 4 shows translations of example dialogues collected in the user study to be described in Section 5 with N being zero (no small talk utterances), one, and three. A longer example can be found in the appendix.

5 User Study

To investigate the effectiveness of the small talk utterances, we conducted a user study.

5.1 Compared Conditions

In this user study, to evaluate the effectiveness of generating small talk utterances, we compared the following three conditions:

NO-STU: The system does not generate any small talk utterances ($N = 0$ in Section 4.6. This is the baseline condition),

1-STU: The system generates one small talk utterance after the user replies to the system question for diet recording ($N = 1$), and

3-STU: The system generates three small talk utterances (three turns) after the user replies to the system question for diet recording ($N = 3$).

We have chosen these for the following reason. First, we conducted a preliminary questionnaire survey to 100 people via crowdsourcing. We showed each participant six sets of dialogues. Each set includes six dialogues each of which has one system question, the user's reply, one of the following, and another system question:

NO**:** nothing,

ONE**:** small talk containing one system turn,

TWO**:** small talk containing two system turns,

THREE**:** small talk containing three system turns,

TWO-CONSECUTIVE**:** one system turn having two consecutive small talk utterances and the user's reaction, and

THREE-CONSECUTIVE**:** one system turn having three consecutive small talk utterances and the user's reaction.

These dialogues were created by the authors based on the functionality of the implemented interview dialogue system. Each participant is asked which he/she likes the best among the six dialogues for each set. Table 3 shows the result. We found the participants liked THREE best.

We also found, however, increasing the number of small talk utterances does not give a better impression to the participants in the trial use of the system. This is probably because the second and third small talk utterances need to react to the user responses to the first small talk utterance and it is difficult to generate utterances appropriate in the context. On the contrary, the dialogues we showed in the above questionnaire survey did not include any inappropriate utterances, thus the participant must have chosen THREE.

ID	Adjective pair		
Q_1	system responses are meaningful	$\leftrightarrow$	system responses are meaningless
Q_2	fun	$\leftrightarrow$	not fun
Q_3	natural	$\leftrightarrow$	unnatural
Q_4	warm	$\leftrightarrow$	cold
Q_5	want to talk to the system again	$\leftrightarrow$	don't want to talk to the system again
Q_6	lively	$\leftrightarrow$	not lively
Q_7	simple	$\leftrightarrow$	complicated

Table 4: Survey items

We therefore used 1-STU in addition to 3-STU in the user study. We also used NO-STU which was dealt with as the baseline.

5.2 Experimental Method

We asked 100 participants recruited through a crowdsourcing site to evaluate the system with different conditions after engaging in the dialogues. We did not collect their personal profiles such as gender and age. The participants accessed the dialogue server to engage in dialogues with the system with the three conditions. The order of the conditions was random. The participants were asked to evaluate the dialogue by rating seven items on a 5-point Likert-scale after finishing the dialogue with the system with each condition. The system finished the dialogue if the number of turns reached 33. For a technical reason, the maximum number of system turns of the dialogue is 34.

When analyzing the evaluation data, we excluded those of eight participants whose dialogues were interrupted due to system problems and who repeated the same utterances many times. The average number of system turns in the dialogues with the NO-STU system, the 1-STU system, and the 3-STU system were respectively 16.5, 23.4, and 30.8.

5.3 Language Understanding Performance

We first evaluated the performance of the language understanding module. We randomly extracted 1,000 user utterances and their understanding results from the collected dialogue logs. We found that the utterance types of the 91.7% of utterances are correctly classified and the semantic contents from the 84.8% of utterances were perfectly extracted.

We also evaluated food group estimation by investigating randomly chosen 200 food group acquisition dialogues. The accuracy of the food group estimation was 84.0%, when we consider the estimation result is correct if one of the candidates the system provided to the user was correct.

5.4 User Impressions

Figure 5 shows the user evaluation results. First, for "simplicity", NO-STU is the best, followed by 1-STU. This is reasonable because the total number of turns becomes smaller when a lower number of small talk utterances are generated.

As for the remaining survey items, we found the 1-STU got significantly higher scores for "fun", "warmth" "want-to-talk-again" and "liveliness" than NO-STU. In addition it is not worse than NO-STU for the other items. This shows small talk utterances improve the impressions of the system.

However, scores of 3-STU are better than those of NO-STU only for "warmth" and "liveliness" and not better than for any items than 1-STU, In addition the "naturalness" scores of 3-STU are significantly lower than those of NO-STU and 1-STU. We discuss this below.

5.5 Discussion

The scores for 3-STU are not good probably because, as we already discussed in Section 5.1, increasing the number of small talk utterances raises the possibility of generating unnatural system utterances. We confirmed this by manually investigating the frequency of inappropriate small talk utterances. We randomly chose five participants and checked their dialogue in 3-STU, and intuitively judged if the small talk utterances are inappropriate considering both the utterance content and the dialogue context. We found that 27 out of 33 first system utterances in the small talk (82%) were appropriate, but that only 18 of 32 (56%) second utterances and 8 of 29 (28%) third utterances were appropriate. We guess this is why 3-STU gives a worse impression to the participants.

We found that the participants are split into two groups depending on the scores for "naturalness" and "liveliness" as shown in Figure 6. Although we have not figured out the exact cause of this, we suspect this is because the expectations of the participants to the ability of the system are different. By looking at the free-form descriptions of impressions of the participants, the participants who scored low in "naturalness" wrote impressions such as "not interesting" and "can't respond well", but the participants who scored high wrote impressions such as "It's fun to think how to speak in order to be understood by the system" and "It's

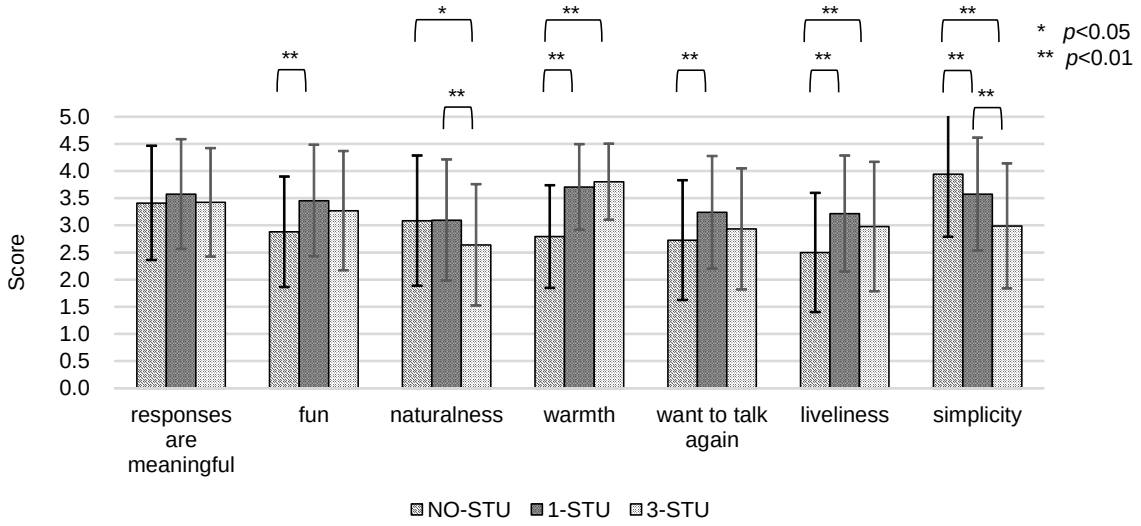

Figure 5: Averages and standard deviations (shown as error bars) of user evaluations on the system. The statistical significances are evaluated using Wilcoxon signed-rank tests.

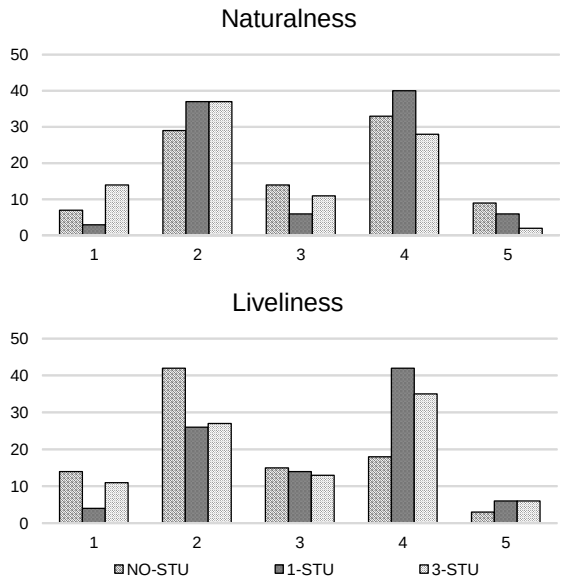

Figure 6: Distributions of "naturalness" and "liveliness" scores

fun to chat with the robot"[2]. That is, those who scored high in "naturalness" did not seem to have high expectations in the ability of the dialogue system. Based on this observation, finding a method for decreasing user expectations is expected to be effective to improve their impressions.

[2]The chat display shows an illustration of a robot.

6 Concluding Remarks

Interviewing is one of the promising applications of dialogue systems technology although not many studies have been conducted so far. This paper proposed to generate small talk utterances to improve user impressions of interview dialogue systems. Based on the proposal, we implemented a Japanese text-based interview dialogue system for diet recording.

The results of a user study showed that small talk utterances give a better impression to users but suggested that generating too many small talk utterances increases the possibility of generating unnatural utterances, making the users' impressions worse.

The user study presented in this paper was based on crowdsourcing. So there can be bias in user attributes such as gender and age. In addition, although our long-term goal is to build interview dialogue systems that users are willing to repeatedly use, the participants used the system only once in the user study. We are planning to conduct another user study to investigate how generating small talk utterances affects the continuous use of the system by recruiting a variety of participants.

The current system uses a fixed number of small talk utterances. We are planning to incorporate a strategy for flexibility selecting utterances from candidates for questions and small talk utterances depending on the context and user reactions. Such a strategy will be learned from the corpus that

we collected in the user study described in Section 5. Furthermore, taking a deep-learning-based approach to utterance selection (Lowe et al., 2015) is one possibility if we can obtain enough training data.

Finally, we plan to investigate how well the results of this study can be applied to interview dialogue systems in other domains.

Acknowledgments

The authors would like to thank Eric Nichols for his valuable comments.

References

Jerome R. Bellegarda. 2013. Spoken language understanding for natural interaction: The Siri experience. In Joseph Mariani, Sophie Rosset, Martine Garnier-Rizet, and Laurence Devillers, editors, *Natural Interaction with Robots, Knowbots and Smartphones*, pages 3–14. Springer.

Timothy Bickmore and Justine Cassell. 2005. Social dialogue with embodied conversational agents. In Jan C. J. van Kuppevelt, Laila Dybkjar, and Niels Ole Bernsen, editors, *Advances in Natural Multimodal Dialogue Systems*, pages 23–54. Springer.

Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2):293–327.

Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. GUS, a frame driven dialog system. *Artificial Intelligence*, 8(2):155–173.

Dan Bohus and Alexander I. Rudnicky. 2009. The RavenClaw dialog management framework: Architecture and systems. *Computer Speech and Language*, 23(3):332–361.

Frederick G. Conrad, Michael F. Schober, Matt Jans, Rachel A. Orlowski, Daniel Nielsen, and Rachel Levenstein. 2015. Comprehension and engagement in survey interviews with virtual agents. *Frontiers in Psychology*, 6. Article 1578.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9.

D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai. 1996. A form-based dialogue manager for spoken language applications. In *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP-96)*, pages 701–704.

Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lefèvre, Patrick Lehnen, Renato De Mori, Alessandro Moschitti, Hermann Ney, and Giuseppe Riccardi. 2011. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Speech and Audio Processing*, 19(6):1569–1583.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING-2014)*, pages 928–939.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*, pages 87–95.

Michael Johnston, Patrick Ehlen, Frederick G. Conrad, Michael F. Schober, Christopher Antoun, Stefanie Fail, Andrew Hupp, Lucas Vickers, Huiying Yan, and Chan Zhang. 2013. Spoken dialog systems for automated survey interviewing. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2013)*, pages 928–939.

Gary Kahn, Steve Nowlan, and John McDermott. 1985. MORE: an intelligent knowledge acquisition tool. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 581–584.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 230–237.

Cheongjae Lee, Sangkeun Jung, Minwoo Jeong, and Gary Geunbae Lee. 2006. Chat and goal-oriented dialog together: A unified example-based architecture for multi-domain dialog management. In *Proceedings of the 2006 IEEE Spoken Language Technology Workshop (SLT-2006)*.

Ryan Lowe, Nissan Pow, Iulian V. Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*, pages 285–294.

Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Mikio Nakano, Atsushi Hoshino, Johane Takeuchi, Yuji Hasegawa, Toyotaka Torii, Kazuhiro Nakadai, Kazuhiko Kato, and Hiroshi Tsujino. 2006. A robot that can engage in both task-oriented and non-task-oriented dialogues. In *Proceedings of the 6th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2006)*, pages 404–411.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Tsugumi Otsuka, Kazunori Komatani, Satoshi Sato, and Mikio Nakano. 2013. Generating more specific questions for acquiring attributes of unknown concepts from users. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2013)*, pages 70–77.

Gabriel Skantze, Samer Al Moubayed, Joakim Gustafson, Jonas Beskow, and Bjorn Granstrom. 2012. Furhat at robotville: A robot head harvesting the thoughts of the public through multi-party dialogue. In *Proceedings of the IVA 2012 Workshop on Real-Time Conversations with Virtual Agents (RCVA 2012)*.

Amanda Stent, Svetlana Stenchikova, and Matthew Marge. 2006. Dialog systems for surveys: the rate-a-course system. In *Proceedings of the 2006 IEEE Spoken Language Technology Workshop (SLT-2006)*, pages 210–213.

David Traum, William Swartout, Stacy Marsella, and Jonathan Gratch. 2005. Fight, flight, or negotiate: Believable strategies for conversing under crisis. In *Proceedings of the International Conference on Intelligent Virtual Agents 2005 (IVA-2005)*, pages 52–64.

Richard S. Wallace. 2008. The anatomy of A.L.I.C.E. In Robert Epstein, Gary Roberts, and Grace Beber, editors, *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, pages 181–210. Springer.

Yorick Wilks, Roberta Catizone, Simon Worgan, Alexiei Dingli, Roger Moore, Debora Field, and Weiwei Cheng. 2011. A prototype for a conversational companion for reminiscing about images. *Computer Speech and Language*, 25(2):140–157.

Appendix: An Example Dialogue of the 1-STU system

Agent	Translation	Original utterance in Japanese*
S:	Let me start an interview. What did you have for breakfast yesterday?	Kore kara intabyû o hajime masu; Kinô no asa wa nani o tabe mashita ka?
U:	I had cereal.	Siriaru desu
S:	I want to eat it, too.	Watashi mo tabe tai desu
U:	You do.	Sô desu ka
S:	What else did you have for breakfast yesterday?	Kinô no chôshoku de hoka ni nani o tabe mashita?
U:	I didn't have any other food.	Tabete nai desu
S:	How much did you have for breakfast yesterday?	Kinô no chôshoku wa dore kurai tabe mashita ka?
U:	I had a normal amount.	Futsûryô desu
S:	How much do you think you had for breakfast yesterday?	Kinô no chôshoku no ryô wa dore kurai datta to omoi masu ka?
U:	A normal amount.	Futsû desu
S:	Sorry, I couldn't understand the answer. Please type in just "yes". [This request is for moving the dialogue forward.]	Sumimasen. Kono ôtô wa ninshiki deki masen deshita; Hai to dake nyûryoku shite kudasai.
U:	Yes.	Hai
S:	What did you have for lunch yesterday?	Kinô no chûshoku, nani o meshiagari mashita?
U:	I had *yakisoba* (Japanese stir-fried noodles).	Yakisoba desu
S:	Is *yakisoba* a one-dish meal?	Yakisoba wa ippin mono desu ka?
U:	Yes.	Hai
S:	Well, you had *yakisoba*.	Hê yakisoba dattan desu ne
U:	Yes, I did.	Sôdesu yo
S:	What else did you have for lunch yesterday?	Kinô no chûshoku de hoka ni meshiagatta mono wa ari masu ka?
U:	Nothing.	Nai desu
S:	How much do you think you had for lunch yesterday?	Kinô no chushoku no ryô wa dorekurai datta to omoi masu ka?
U:	A proper amount.	Teki ryô desu
S:	Can you tell me what you had for supper yesterday?	Kinô no yûshoku ni tabeta mono o oshiete kudasai
U:	I had curry and rice	Karê desu
S:	I want to eat it!	Tabe tai desu!
U:	I see!	Sokka!
S:	What else did you have for supper yesterday?	Kinô no yûshoku de karê igai ni meshiagatta mono wa ari masu ka?
U:	Nothing.	Naidesu
S:	Did you have a lot for supper yesterday?	Kinô no yûshoku wa gattsuri tabe mashita ka?
U:	A lot. [Semantic content extraction failed.]	Gattsuri desu
S:	How much did you have?	Dorekurai tabemashita ka?
U:	A lot.	Gattsuri desu
S:	This interview is finished. Thank you.	Kono intabyû wa shûryô desu. Arigatô gozaimasu

* Note that the user and the system used Japanese characters for interaction.