Learning Dialogue Strategies for Interview Dialogue Systems That Can Engage in Small Talk

Tomoaki Nakamura, Takahiro Kobori, and Mikio Nakano

Abstract This paper proposes a method with which an interview dialogue system can learn user-friendly dialogue strategies. Conventional interview dialogue systems mainly focus on collecting the user's information and simply repeat questions. We have previously proposed a method for improving user impressions by engaging in small talk during interviews that performs frame-based dialogue management and generates small-talk utterances after the user answers the systems questions. However, the utterance selection strategy in the method was fixed, making it difficult to give users a good impression of the system. This paper proposes a method for learning strategies for selecting system utterances based on a corpus of dialogues between human users and a text-based interview dialogue system in which each system utterance was evaluated by human annotators. This paper also reports the results of a user study that compared the proposed method with fixed utterance selection strategies.

1 Introduction

We aim to develop interview dialogue systems that ask questions and obtain a user's information. One example is a dialogue system for diet recording that asks what the

Tomoaki Nakamura
The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, Japan
e-mail: tnakamura@uec.ac.jp

Takahiro Kobori
The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, Japan
e-mail: tkobori@yahoo-corp.jp
(Present affiliation: Yahoo Japan Corporation, 1-3 Kioicho, Chiyoda-ku, Tokyo, Japan)

Mikio Nakano
Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama, Japan
e-mail: nakano@jp.honda-ri.com

user eats or drinks. Workforce reduction is one potential advantage of such dialogue systems. Humans are more likely to disclose their information to dialogue systems than human interviewers [6], and thus may be able to obtain more information than human interviewers do.

Only a few studies of dialogue systems have delved into their application to interviews. For example, Stent et al. built a dialogue system for questionnaires that was applied to university course evaluations [9] and Johnston et al. developed a dialogue system for social surveys [3]. These studies mainly focused on collecting user information. Although the users might use the dialogue systems once, we believe they would be unwilling to use them continuously. In contrast, non-task-oriented dialogue systems [10, 11, 2], designed for users to enjoy chatting with, have been widely studied.

We previously proposed a dialogue management method that mainly engages in an interview and sometimes engages in small talk, which improves users impression of the interview system [4]. We implemented a system that generates small-talk utterances during an interview based on heuristic rules. A user study revealed that the small talk improves the impressions of the system, but we found that it was difficult to always select appropriate utterances with static heuristics.

In this paper, we propose to introduce utterance selection rules learned from annotated dialogue corpus and evaluate the effectiveness of the proposed system.

2 Proposed Method

The proposed method selects system utterances from questions for interviewing and small-talk utterance candidates as was done in our previous method [4]. The difference is that it takes a learning-based approach for selecting system utterances. In our proposed method, evaluation score y_c that indicates the appropriateness of the c-th candidate u_c is estimated from dialogue history x using regression function f:

$$y_c = f(\phi(u_c; x)). \tag{1}$$

A candidate utterance is selected based on y_c. $\phi(u_c; x)$ denotes feature vectors extracted from system utterance candidate u_c and dialogue history x, using the following features:

- Bag of words (BoW) of the user's last utterance,
- The number of turns from the beginning of the dialogue,
- The number of small-talk utterances generated up to that point in time,
- BoW of the system and user's utterances during the last three exchanges (three system and three user utterances),
- Words that occur in both the system utterance candidate and the user's last utterance, and
- The sequence of the type of the last system utterance and the type of the candidate utterance.

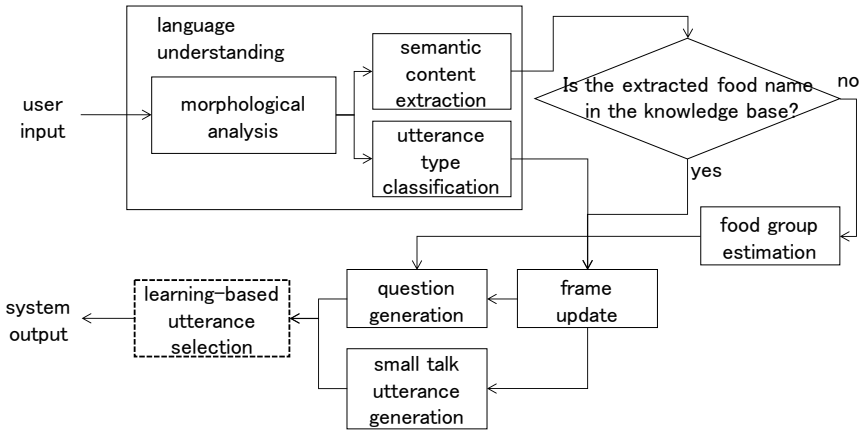

Fig. 1 Architecture of the proposed dialogue system. The dashed rectangle shows the novel module being added to the previous system.

The regression function parameters are learned from a corpus, in which system utterances in the dialogue logs are manually evaluated using 5-point Likert scale. By estimating this evaluation score, the system can select more appropriate utterances.

3 An Interview Dialogue System for Diet Recording

This section describes a system that implements the proposed method. It can engage in interview dialogue for diet recording. Figure 1 depicts its architecture. The system is basically same as the system previously presented in [4], but there are the following differences:

- The training data for understanding the user's utterances are updated by adding sentence templates.
- The content of the knowledge base is significantly larger.
- A new method for selecting system utterances is introduced.

3.1 Understanding User Utterances

User utterances are classified into three types: greetings, affirmative utterances (including replies to system questions), or negative utterances. We chose these types because they allow the system judge whether frames should be updated based on the user's utterance. A user utterance to tell that he/she had some food or drink

is classified as affirmative, and the frame is updated according to its content. We used logistic regression (LR) for utterance type classification, with a bag-of-words (BoW) of the user's utterances serving as features. In this study, we used Mecab [5] for morphological analysis and Liblinear [1] for implementation of LR.

Furthermore, from the user's utterance, the system extracts five semantic contents–food and drink, ingredients, food group, amount of food, and time at which the food or drink was consumed–using conditional random fields (CRFs). We used the unigrams and bigrams of the surface form, original form, and part of speech of the words as CRF features, and CRFsuite [7] for implementation of the CRFs. Moreover, in addition to CRFs, semantic content is extracted by a rule-based method that utilizes regular expressions, which makes it possible to accurately extract predefined words and extract words difficult to extract only with CRFs.

The implementation of this module is the same as that of our previous system [4]; however, the training dataset was updated. It now includes 3,659 sentences generated by replacing the content words in 603 sentence templates for training the LR and CRFs.

3.2 Dialogue Management for Interviews

As in our previous work [4], we utilize frame-based dialogue management. The slots of the frame are meal time (breakfast, lunch, and dinner), composition, and amount (Fig. 2). The frame is updated according to user utterances. To fill in slots appropriately, the system must know the food groups of the extracted food names. For example, if the user says, "I had ramen for lunch," the system must understand that "ramen" is a one-dish meal and fill "ramen" into the "one-dish-meal" slot. The food group for a food names is determined using the knowledge base, examples of which are shown in Table 1. In this implementation, 2,134 food and drink instances are included in the knowledge base, as compared to only 304 instances in the previous implementation [4].

However, all food names are not covered by the knowledge base, and the user might tell a food that is not included. In such a case, we employ LR-based food group estimation as proposed by Otsuka et al. [8] with BoW, unigrams and bigrams of characters, and character type (hiragana, katakana, Chinese characters, and Latin alphabet) as features.

The system asks a question to fill vacant in slots by referencing the current frame, or chats using small talk. The question utterances are not fixed expressions, and candidates that include various expressions are generated. Finally, the system selects one candidate as a system utterance.

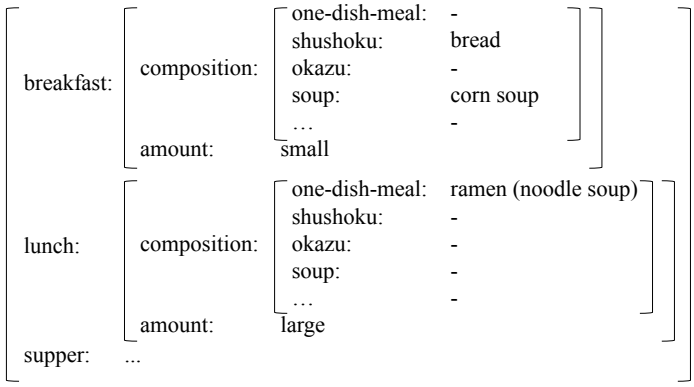

Fig. 2 Example frame.

Table 1 Knowledge base content.

Food group	Example instances	#
shushoku (side dish mainly containing carbohydrates)	steamed rice, bread, cereal	152
okazu (main or side dish containing few carbohydrates)	hamburg steak, fried shrimp, grilled fish	668
soup	corn soup, miso soup	70
one-dish meal	sandwich, noodle soup, pasta, rice bowl	695
drink	orange juice, coffee	343
dessert	cake, pancake, jelly	134
confectionery	chocolate, donut	72
total		2,134

3.3 Small Talk Generation

Small talk candidates are generated as system utterances depending on the user utterance type and the context as in the previous work [4]. If a user utterance is estimated to be affirmative, all utterances except for negative utterances are generated as response candidates; for example, utterances such as, "I like that too!" (self-disclosure); "That's great!" (empathy); and "Was it tasty?" (question) are generated. However, if the user's utterance is estimated to be negative, response utterances are generated as candidates: for example, "That's too bad$\cdots$" and "You should have something you like instead." If specific foods are included in the user's utterance, corresponding small talk is included in the response candidates. For example, if the user's utterance includes "cookie," then "I heard it's originally from Persia," is generated. In addition, if the user's utterance includes the amount of the meal, small talk about this becomes candidates for reply. For example, if the user says "I didn't eat much," a system utterance such as "It's better to eat properly for your health," is generated as candidates.

To collect small talk, we conducted a crowdsourced questionnaire, from which we obtained 442 utterances. In the previous implementation, we used all 442 utter-

Table 2 Small talk utterances obtained via questionnaire.

Type	#
showing empathy	24
commenting that the expressed amount is large	15
commenting that the expressed amount is small	43
asking a question	5
self-disclosure	2
backchannel	6
giving an impression of the user's negative answer	10
reaction to individual food	302
Total	407

ances; however, some of these were inappropriate. We removed these instances, and 407 utterances are used in the current system, details of which are shown in Table 2.

3.4 Selection of System Utterances

As previously described, once questions and small talk candidates are generated, the system must select one as an utterance. In our previous study, an utterance was selected based on heuristic rules; the system chose to ask a question or make small talk utterance based on the rule where the number of consecutive small talk utterances is fixed, and one candidate in the selected type (question or small talk) was selected randomly. However, this fixed strategy often gave users a bad impression of the system. To solve this problem, we utilize a learning-based approach to select the system utterances using the proposed method explained in Section 2.

First, evaluation scores for all candidates are estimated, and then the system selects whether a question or small-talk utterance should be generated. Here, we consider it better for the system to generate small-talk utterances preferentially to facilitate smooth dialogue; therefore, if there are small-talk utterances whose evaluation scores are greater than threshold T, a small-talk utterance is selected. Otherwise the question with the highest evaluation score selected as the system utterance. When a small-talk utterances is to be selected, then one with a score greater than T is selected based on the following conditions:

1. Utterances that have already used in a dialogue cannot be selected again,
2. Candidate with the first, second, and third highest score is reserved, and the fourth highest candidate is selected, and
3. If there are fewer than four candidates, the one with highest score of the reserved candidates is selected

Condition 1 prevents unnaturalness in the dialogue. Regarding Condition 2, we found that small-talk utterances with higher scores are likely to be general, and can be used as responses to various user utterances (e.g, "Was it tasty?"). In contrast,

utterances for specific foods and drinks are likely to have lower scores. Therefore, if the utterance with the highest score is always selected, the dialogue becomes boring. Condition 2 prevents such a situation, as small-talk utterances with lower scores (though still higher than T), which are, to an extent, more natural, are selected as system utterances. Furthermore, if it is difficult for the system to respond to the user's utterance, Condition 3 prevents a breakdown of the dialogue by using reserved general utterances. This study employed an extremely randomized tree regressor (ETR) to estimate the score, and two ETR models are separately learned for selecting questions and small talk.

4 Experiments

We conducted a user study to investigate the effectiveness of our proposed method.

4.1 Corpus for Training System Utterance Selection

To train the ETRs, we constructed dialogue corpus with evaluation scores. The corpus included 4,523 system utterance candidates randomly chosen from dialogue logs obtained in the previous study [4]. We measured annotation agreement between two annotators using 2,018 system utterance candidates. The weighted kappa statistics, which measures the agreement between the scores of two annotators, and it was 0.452. Since this value represents moderate agreements, we used the annotations of one of the annotators.

4.2 Experimental Setup

We compared four systems: an interview dialogue system without small talk (NO-STU); a system that generates one small-talk utterance after each user utterance (1-STU); and two systems that generate small-talk utterances using ETR models (ETR and ETR*), whose thresholds T, explained in Sec. 3.4, are 2.3 and 2.8, respectively. To evaluate the systems, 160 people participated, and 40 people interacted with each system and evaluated it. We assigned participants to each system so that ages and genders of people using each system became roughly equal. If the purpose of the experiment is to make significant differences among the systems clear, it might be better that same subjects use different systems. However, we would like to find how the impression against the same system changes depending on the number of times of its use in this experiment and, therefore, we assigned the subjects as described above. The subjects accessed the server on which the dialogue systems were running and engaged in the interview dialogue. In this experiment, to investigate how the

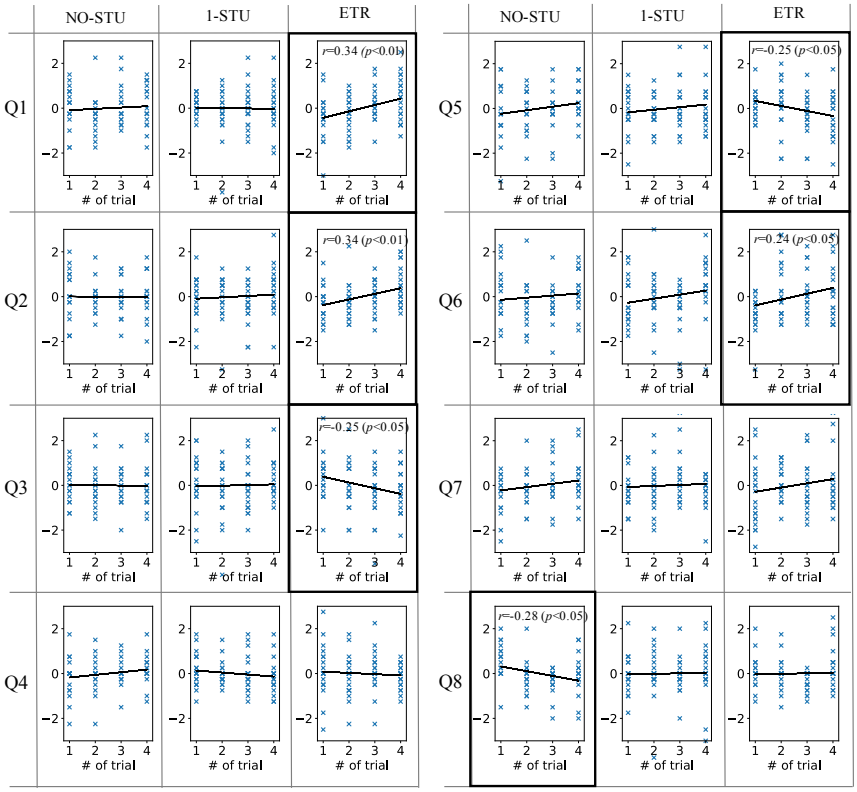

Fig. 3 User evaluations of each dialogue system. Each point represents a user evaluation; solid lines are the regression line of all user evaluations. Graphs in a black rectangle denote significant items ($p < 0.05$).

impressions of each system changed with continuous use, the subjects were required to use the assigned system at least four times in a week and up to once per day. After dialogue with the systems, we asked the participants to evaluate the dialogue using a five-point Likert scale for 8 survey items, shown in Table 3, and describe their impression of the dialogue. Considering the participants load, the dialogue was automatically stopped after 40 turns. Some participants who did not follow these experimental conditions or could not use the system because of system errors were excluded. Moreover, we found that a threshold of 2.8 for ETR* was too high, as there were few small-talk utterances candidates whose score was higher than 2.8 toward the end of each dialogue. We consider this is not a system that we intended in this paper, and this system is excluded for the evaluation. Therefore, the number of participants whose evaluations were used in this investigation was 57 (NO-STU:17, 1-STU:20, ETR:20).

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 continue to talk	$\leftrightarrow$ don't want to continue to talk
Q_6	lively	$\leftrightarrow$ not lively
Q_7	simple	$\leftrightarrow$ complicated
Q_8	want to talk to the system again	$\leftrightarrow$ don't want to talk to the system again

Table 3 Survey items.

4.3 Questionnaire Evaluation

In this experiment, we analyzed how the users impression of the system in each condition changed during four dialogues. To make this change clearer, evaluation score e_{in} of user i against the n-th dialogue was normalized to an average of zero:

$$\hat{e}_{in} = e_{in} - \frac{1}{4}\sum_{n} e_{in}. \tag{2}$$

The normalized evaluation scores of all subjects are plotted in Fig. 3. Each graph shows each survey item: the horizontal axis of each graph represents the number of dialogues, and the vertical axis represents the normalized evaluation scores. We conducted a regression analysis in which the number of dialogues is the explanatory variable and the evaluation score is the dependent variable. Graphs with black rectangles denote significant items using a level of significance of 0.05.

There were significant positive correlations in Q1, Q2, and Q6 of the ETR condition; users felt that system responses became more meaningful and that the dialogues became more fun and lively as the number of dialogues increased. In contrast, in other conditions, significant correlation was not seen in these survey items. The system in the NO-STU condition simply repeated questions; participants who used the NO-STU system described their impression with phrases such as "It utters the same pattern every time," and "Repeating the same questions is boring." Such impressions decreased in conditions using small-talk utterances. In the 1-STU condition, another reason for giving a bad impression is that the system ignored the user's response to its small talk utterance and instead asked the next question. In the ETR condition, however, small-talk utterances can be generated continuously, which was more fun for users.

However, there were negative correlations in Q3 and Q5 of ETR condition; users felt that system utterances were unnatural, and as the number of dialogues increased, they did not want to continue talking with the system. We believe this result was caused by expectations for the system. In the ETR condition, the system generated various small-talk utterances, and the subjects expected that the system were able to talk flexibly. However, the current system cannot engage in flexible conversa-

tion, creating a gap between users expectations and the current systems abilities and decreasing users evaluations.

One of the most important results is the significant negative correlation in Q8 of the NO-STU condition; users were bored and did not want to talk with the system without small-talk utterances again. Even the 1-STU system, which generates a small-talk utterance after a user utterance, prevented this negative effect. However, positive correlation in Q8 of the ETR condition, which we expected, was not seen. We plan to investigate the factors that are important in making the user willing to talk to a system again.

5 Conclusion

We have developed a dialogue system that generates small-talk utterances to improve user-friendliness and encourage future use. In this study, we proposed a method for system utterance selection based on an evaluation score estimated by a regression model. We conducted a user study to investigate the effectiveness of the proposed method and small-talk utterances, and found that participants found dialogue to be more fun and lively when generated by the proposed method than by the previous system, which used heuristic rules. Moreover, we confirmed that fewer participants wanted to talk to a system that did not generate small talk again as the number of dialogues increased. This result indicates that small-talk utterances are important for continuous use of a dialogue system. However, the proposed system received poor evaluation in some survey items, so we plan to investigate the reasons for that.

References

1. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9**, 1871–1874 (2008)
2. Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing. In: International Conference on Computational Linguistics, pp. 928–939 (2014)
3. Johnston, M., Ehlen, P., Conrad, F.G., Schober, M.F., Antoun, C., Fail, S., Hupp, A., Vickers, L., Yan, H., Zhang, C.: Spoken dialog systems for automated survey interviewing. In: Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 329–333 (2013)
4. Kobori, T., Nakano, M., Nakamura, T.: Small talk improves user impressions of interview dialogue systems. In: Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 370–380 (2016)
5. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis. In: the Conference on Empirical Methods in Natural Language Processing, pp. 230–237 (2004)
6. Lucas, G.M., Gratch, J., King, A., Morency, L.P.: Its only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior **37**, 94–100 (2014)

7. Okazaki, N.: CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/ (2007)
8. Otsuka, T., Komatani, K., Sato, S., Nakano, M.: Generating more specific questions for acquiring attributes of unknown concepts from users. In: Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 70–77 (2013)
9. Stent, A., Stenchikova, S., Marge, M.: Dialog systems for surveys: the rate-a-course system. In: IEEE Spoken Language Technology Workshop, pp. 210–213 (2006)
10. Wallace, R.S.: The anatomy of alice. Parsing the Turing Test pp. 181–210 (2009)
11. Wilks, Y., Catizone, R., Worgan, S., Dingli, A., Moore, R., Field, D., Cheng, W.: A prototype for a conversational companion for reminiscing about images. Computer Speech & Language **25**(2), 140–157 (2011)