Analyses of Example Sentences Collected by Conversation for Example-Based Non-Task-Oriented Dialog System

Yukiko Kageyama, Yuya Chiba, Takashi Nose, and Akinori Ito, *Member, IAENG*

***Abstract*—Designing an example database is important for handling various users' utterances in an example-based dialog system, and several approaches to constructing the database have been proposed. This paper focuses on a method for collecting the example sentences through actual conversations with the system. Several studies employ this approach for constructing the dialog system, but conventional research lacks attentive analyses. In this study, we analyzed how many examples can be collected from the interactions, and investigated the characteristics of the collected examples. The experimental results show that the response accuracy improved with the increase in number of the interactions, and the examined collection method is effective for collecting examples of consecutive utterances. In addition, subjective evaluation comparing the databases constructed using actual conversation and the fully-handcrafted databases was conducted through dialog experiments. The results showed that the examined approach can obtain higher subjective scores than the comparative approach in terms of user satisfaction, dialog engagement, intelligence, and intention of talking.**

***Index Terms*—example collection, actual conversation, spoken dialog system, example-based dialog system**

I. Introduction

FOCUS has been placed on non-task-oriented spoken dialog systems [1–3] in contrast to traditional task-oriented dialog systems, such as train information services [4] or information recommendation [5]. The non-task-oriented dialog system aims for a short-conversation with the user rather than achieving a specific task goal, but it is reported that a conversational agent capable of chat-like talk can improve the performance of the user even in a task-specific dialog system [6]. Therefore, the non-task-oriented dialog appears to be increasingly important for future spoken dialog systems.

General task-oriented dialog systems are built to have an internal state. The system updates the state depending on the user's utterance to make an appropriate response. On the other hand, many non-task-oriented dialog systems are example-based systems because defining the state is difficult for such dialog systems. The example-based system generates a response as a prepared sentence corresponding to the most similar example to the user's utterance. Therefore, the quality of the example-response database directly affects the performance of the dialog system. In particular, a developer needs to meet the following two requirements:

Manuscript received August 9, 2017; revised December 9, 2017. Funding was provided by Grants-in-Aid for Scientific Research (Grant No. JP15H02720 and JP17H00823).

Y. Kageyama, Y. Chiba, T. Nose, and A. Ito are with the Graduate School of Engineering, Tohoku University, 6-6-05, Aramaki Aza Aoba Aoba-ku, Sendai, Miyagi, Japan (e-mail: {yukiko.kageyama.s2@dc, yuya@spcom.ecei, tnose@m, aito@spcom.ecei}.tohoku.ac.jp)

1) Construction of an example database coinciding with the target dialog domain to handle any user's utterance.
2) Creation of natural responses corresponding to each example.

This study mainly focuses on requirement 1) because preparing the various examples is important for the non-task-oriented dialog to respond to the utterances of highly diverse user.

Numerous methods for constructing the example database have been studied so far. They are roughly classified into two categories; manual development and collection from web resources. In the manual method, the quality of the database is assured because humans judge the naturalness of the example-response pairs, but the developer must create the database by imagining the possible conversations. Therefore, it is difficult to create examples of utterances that depend on the context of the dialog, such as answer utterances from the user responding to the system's utterance. Database construction based on automatic collection can accumulate a large scale of example-response pairs, but it is also hard to prepare the pairs considering the context of the dialog. Another problem with the automatic approach is that text-based interaction is greatly different from speech that occurs in the actual conversation.

Because most utterances that occur in every day chat-like talk depend on the context, it is important to collect examples corresponding to the user's utterance depending on the context such as consecutive utterances from the first interchange. One of the methods that can collect examples coinciding with both the speech communication and the dialog context is using the user's utterances during actual conversation with the system. This approach is expected to not only collect examples observed in a real situation, but also effectively collect examples for consecutive utterances if the database is updated dialog by dialog. Several studies have actually taken this approach for constructing the dialog system (e.g., [7, 8]). However, since they employ a framework whereby additional example-response pairs are appended after relatively long-term interactions, we do not know how many examples are needed to cover the context-dependent utterances. In addition, it is not clear if the developed database really surpasses the handcrafted one in terms of user experience of coincidence with the actual speech interaction.

Therefore, this paper investigates the characteristics of the database collected from actual conversations compared with the handcrafted database, by analyzing 1) the relationship between the number of interaction and the appropriateness of the system's responses, 2) the characteristics of the collected examples, and 3) the subjective evaluation scores of the

dialog experiments.

This paper is organized as follows. The following section describes the conventional approaches used for example collection. Section III explains the dialog system constructed for the experiments. Section IV describes the procedure for collecting examples based on actual conversation and investigates the appropriateness of the responses with respect to the number of interactions. Section V describes the dialog experiment using the developed database and a handcrafted database for subjective evaluation, and several quantitative analyses are conducted. The conclusions are presented in Section VI.

II. Conventional approaches of example collection

Research on the non-task-oriented dialog system tracks dates back to Eliza [9] or A.L.I.C.E [10]. Several approaches have been studied for the response selection of the non-task-oriented dialog system, and the dialog system based on the example-response database is widely employed today [11]. In the most typical approach, experts compose the example and response sentences (we call this "handcrafting"). For example, Dickerson et al. developed a diagnosis training system that used question-answer pairs to simulate the interaction between a doctor and patients, where the pairs were created by experts [12]. Virtual agents developed by Kenny et al. simulated patients with post traumatic stress disorder (PTSD) by using question-response pairs manually mapped by a domain expert [13]. In addition to the medical field, Nisimura et al. introduced an example-based system for tourist guidance [7], and Bobrow et al. used the sentence fragments collected from human conversation for tour guides [14]. These systems use pairs created by experts in the dialog task domain. The example-response pairs created by experts seem to be efficient in such domain-specific systems because they are capable of assuming the actual conversation. On the other hand, creating the example-response pairs for a non-task-oriented dialog is not easy. Sugiyama et al. studied an approach for creating question-response pairs by crowdsourcing and analyzed the tendency of the questions [15]. The analysis results revealed that it was difficult to collect context dependent questions, which require knowledge of the previous utterances. Coverage of such questions greatly affects the naturalness of chat-like talk, and so we need to develop a method to collect as many of such questions as possible.

Another approach is to construct the database using web resources [16, 17], and this method is expected to substitute manual database development and reduce the construction cost. For example, Chung et al. constructed a weather forecast system by using web resources [18] and Katz utilized a knowledge base constructed from web data for the dialog [19]. In addition, an approach based on the Wizard of Oz method using a dialog simulator has also been examined [20]. The dialog systems in these research studies were task-oriented example-based systems. On the other hand, a TV show's dialog scripts [21] or movie scripts [22] are also utilized as a source for non-task-oriented systems, but it is considered difficult to use dialogs conducted in a specific situation as the source for general non-task oriented dialog system.

Besides the simple methods based on the example-response database, statistical approaches such as neural-network based conversational modeling have been studied actively [23–25]. These systems do not use prepared example-response pairs but instead train the model to generate a response sentence based on the user's input utterance. The naturalness of the response sentences can be improved by using large-scale training data, but this has not yet reached the level of human-generated response sentences.

In contrast to the above-mentioned approaches, example collection by actual interaction has the following characteristics:

- We can collect frequent examples because the utterances that occur in non-task-oriented dialog is assumed to be biased if the topic is limited.
- Examples of consecutive utterance not included in the database can be collected effectively if the dialog progresses properly.
- It is easy to create natural responses because the human developer composes responses corresponding to the collected examples.

This approach was examined as Human-centered Distributed Conversational Modeling (HDCM) [26]. The study reported that the response accuracy tends to improve by adding the example-response pairs step by step. However, HDCM employs keyboard input, and it is not clear whether keyboard-based examples are useful for spoken dialog systems. In addition, the study lacks the fundamental analysis that is our focus, such as how many examples can be collected by iterating interactions or the characteristics of the collected examples.

III. Dialog system for the experiment

Fig. 1 demonstrates the flow of the non-task-oriented dialog assumed in this paper. Dialog between a user and the system is called *an interaction*. One interaction consists of the interchange of an utterance by the user and a response by the system. Different users talk with the system interaction by interaction, and the index of the user is denoted as i. In addition, the database is updated after every interaction for efficient collection of contextual examples.

A. Dialog management of example-based system

An experimental dialog system was constructed based on the example-based approach. The system calculates the similarities between the user's utterances obtained by speech recognition and example sentences in the database, and then selects a response corresponding to the most similar example. In this study, cosine similarity is used for the similarity calculation. Let $\boldsymbol{q}$ the word vector of the user's utterance, $\boldsymbol{d}$ the word vector of an example sentence, and q_j, d_j be the j-th element of $\boldsymbol{q}$, $\boldsymbol{d}$. The cosine similarity is represented as follows:

$$\cos(\boldsymbol{q}, \boldsymbol{d}) = \frac{\boldsymbol{q} \cdot \boldsymbol{d}}{|\boldsymbol{q}||\boldsymbol{d}|} = \frac{\sum_j q_j d_j}{\sqrt{\sum_j q_j^2}\sqrt{\sum_j d_j^2}} \tag{1}$$

Then, the most similar example $\hat{\boldsymbol{d}}$ is selected as:

$$\hat{\boldsymbol{d}} = \arg\max_{\boldsymbol{d}} \cos(\boldsymbol{q}, \boldsymbol{d}) \tag{2}$$

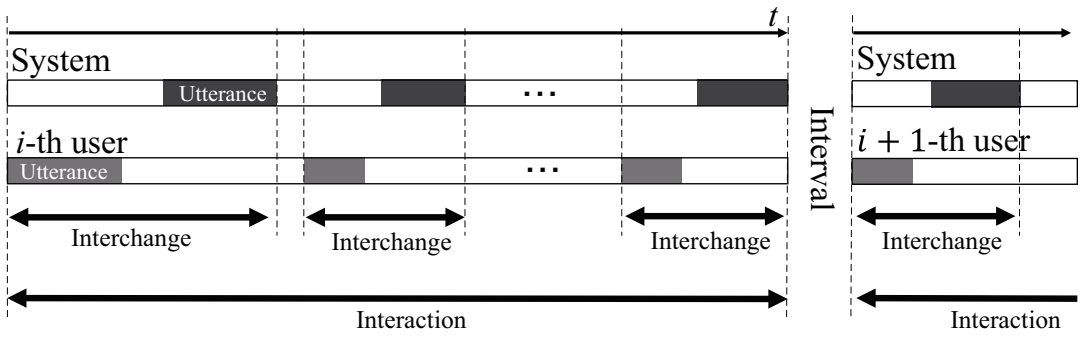

Fig. 1. Structure of non-task-oriented dialog in this study

Fig. 2. Dialog agent

TABLE I
NUMBER OF PAIRS IN INITIAL DATABASE

Category	Cooking	Movies	Meal	Shopping
Greetings	47	47	47	47
Backchannels	6	6	6	6
Topics	369	376	434	469
Total	422	429	487	522

TABLE II
EXAMPLE-RESPONSE PAIRS IN INITIAL DATABASE (TRANSLATION FROM JAPANESE)

Category	Example	Response
Greeting	Hello.	Hello.
Cooking	What did you cook?	I cooked *nikujaga*.
Movies	What is your favorite movie?	*Titanic*.
Meals	What did you have for dinner?	I had *sushi*.
Shopping	What kind of book do you like?	I don't have a particular favorite genre, but I often read *Keigo Higashino*.

The system sends the response sentence corresponding to $\hat{d}$ to the speech synthesis module, and the system's speech is generated.

We implemented the system based on MMDAgent version 1.5 [27], which is an open-source toolkit for building the speech interaction system. A developer can construct the dialog system by integrating the speech recognition module, speech synthesis module, and 3DCG rendering module for a virtual agent. MMDAgent employs Julius as the speech recognizer [28] and Open JTalk as the HMM-based speech synthesizer [29]. Dialog management is based on a finite-state transducer (FST), that allows the developer to describe flexible dialog transition, and which we applied to the example-based dialog system. In our experiments, we employed an embodied female agent to reduce the user's mental load. Fig. 2 shows the appearance of the dialog agent.

B. Non-task-oriented dialog with topic dependent database

The target dialog domain was chat-like talk between friends, which is one type of non-task-oriented dialog. In the experiments, the participants were instructed to ask the agent what she did yesterday on the assumption that she led a human-like life. The example-response databases for the first dialog are required for dialog collection. These databases are called initial database in this paper. We prepared the initial databases for four topics to compare the topic dependency of the collected examples. The topics were cooking, movies, meals, and shopping. The initial databases were composed of example-response pairs corresponding to the greetings, backchannels, and topic-specific interchange. These databases were manually created, and were constructed with questions about the agent's assumed daily events. Each database had the same example-response pairs for greetings and backchannels. The example-response pairs in the initial databases were created by one person (the first author) for consistency. Table. I shows the number of pairs in the initial databases, and Table. II shows the example-response pairs included in each category.

C. Conditions of speech recognition

In the speech recognition module, an acoustic model distributed with MMDAgent was used for the decoding. This model was trained using 44,556 utterances (145 male speakers, 144 female speakers) from the Japanese Newspaper Article Sentences (JNAS) database [30]. The features for training were 12-dimensional Mel-frequency cepstral coefficients (MFCC) including Δ coefficients and the Δ coefficient of the normalized energy. The phone model was a three-state left-to-right GMM-HMM of the triphone model. The number of the mixtures of Gaussian was 16. A tri-gram language model was trained using the transcriptions of academic presentations and simulated public speech in the Corpus of Spontaneous Japanese (CSJ) database [31] and the example

Algorithm 1 Example collection by conversation

Set initial database D_1^k
for $i = 1, ..., I$ **do**
  i-th user talks with the system $S(D_i^k, LM_i^k)$
  Get example sentences E_i^k from conversation log
  Update the database $D_{i+1}^k \leftarrow D_i^k \cup E_i^k$
  Train language model LM_{i+1}^k by $D_{csj} \cup D_{i+1}^k$
end for

sentences of the initial databases to accommodate task-specific utterance. The language model was retrained after every interaction by incorporating the example sentences obtained from the interaction.

D. Conditions of speech synthesis

This study used a conversational-style speech corpus uttered by an amateur female speaker for the synthesis of spontaneous speech. We used 503 phonetically balanced sentences from the ATR Japanese speech database [32], 150 sentences for which the ending was transformed to a conversational-style, 150 conversational-style interrogative sentences, and 250 emotion-specific sentences to train an acoustic model for the speech synthesis. This corpus includes speech samples of four styles (happy, angry, sad, and neutral) and we used the "happy" style speech. The STRAIGHT [33] was used to extract the acoustic features including a 60 dimensional Mel-cepstrum with 0th coefficient, and their Δ and Δ^2 coefficients for training the acoustic model. The five-state left-to-right hidden semi-Markov models (HSMMs) [34] were trained for the speech synthesis. The output probability distribution function was modeled by a single Gaussian with a diagonal covariance matrix.

IV. Example collection by actual conversation

A. Procedure for example collection

Algorithm 1 shows the procedure for the collection of examples for topic k. Let I be the total number of interactions, D_i^k the database of topic k at the i-th interaction, and D_{csj} the document set of the CSJ. LM_i^k represents the language model for topic k at the i-th interaction, and E_i^k represents the example sentences for topic k obtained from the i-th interaction. $S(D_i^k, LM_i^k)$ represents the dialog system using D_i^k as the example-response database and LM_i^k as the language model. In the experiment, the index of the iteration corresponds to the user index, and each user talks with the system only once during the dialog collection.

Automatic extraction of the example sentences E_i^k from the conversation log and response creation are required to completely automate the procedure. However, the focus of this paper is investigating the usability of the examples collection by conversation, and both were conducted manually. In addition, we used the transcriptions of utterances not included in D_i^k as E_i^k.

B. Conditions of example sentence collection

The dialog experiments for collecting examples were conducted in a soundproof chamber. Twenty-five persons (15 males, 10 females) participated in the experiments. Firstly, the participants were presented with the list of topics, and were instructed to ask what the agent did yesterday. Here, the dialog example shown in Table. III was also presented to the participants. The participants were instructed to talk with the system about the provided topic, and utterances relating to self-disclosure were permitted. The topics of the dialog were presented to the participants in random order. The database and the language model were switched each time the topic changed. The system assumes that the user talks along the topic as shown in the table. If the user makes an unanticipated utterance, the story breaks because the system always chooses its response using only the current utterance. We instructed the participant to talk with the system for the specified period, regardless of whether the story continues or not. Table. IV shows an example of a dialog that occurred during the dialog collection.

TABLE III
Dialog example for instruction (translation from Japanese)

User	What did you do yesterday?
System	I studied at home.
User	What subject did you study?

TABLE IV
Example of collected dialog (translation from Japanese)

User	What did you do yesterday?
System	I cooked dinner at home.
User	What did you cook?
System	I cooked *nikujaga*.
User	Sounds good.
System	I know, right?
User	Did you eat alone?
System	Yes. I ate alone while watching TV.

The experiment was separated into two sections. In the first section, the participants made ten interchanges to investigate the appropriateness of each response of the current database D_i^k. The participants evaluated each response as "appropriate" or "not appropriate." After that, as the second section, they engaged in three minutes of dialog for the collection of examples without the appropriateness evaluation.

We also investigated the appropriateness of the response to the transcription of the user's utterance because a speech recognition error can sometimes cause the generation of an incorrect response. The response to the transcription was selected based on Eq. (1). The appropriateness of the response was decided by a majority-vote of three evaluators (1 male, 2 females).

C. Measurement of appropriateness

We defined the coverage of the user's utterances as the measurement of the appropriateness of the system's responses. The coverage of topic k of the i-th interaction is denoted as C_i^k and calculated as follows:

$$C_i^k = \frac{R_i^k}{N_i^k} \tag{3}$$

R_i^k is the number of responses evaluated as "appropriate" and N_i^k is the number of total responses of topic k of the i-th interaction. The appropriateness evaluation was conducted only for the first ten interchanges as mentioned in Section IV-B, and thus $N_i^k = 10$. In addition, the coverage of the i-th

interaction C_i is calculated as the average of the coverage in the topics.

$$C_i = \frac{1}{K} \sum_{k=1}^{K} \frac{R_i^k}{N_i^k} \quad (4)$$

Here, K is the total number of topics ($K = 4$). In the following section, the results of automatic speech recognition (ASR) are denoted as RECOG and that of the transcription are denoted as TRANS in the figures and the tables.

D. Analysis of the relationship between response appropriateness and number of interactions

Using the response appropriateness measure defined in Section IV-C, we analyzed the experimental results. Fig. 3 represents the word error rate (WER) for the speech recognition results for each interaction. The average WER was 24.10%. As shown in the figure, the overall performance of the speech recognition tends to improve by iterating the interaction although the WER for the 17th and 18th participants were relatively high. These results were due to the incremental adaptation of the language model.

Fig. 4 shows the trend of the coverage C_i. In the figure, we calculated the average coverage score for every five interactions. The error-bar is the standard error. The solid line shows the trend of the coverage for the ASR results. As shown in the figure, the coverage improves until the 11–15th interactions, and remains flat after that. On the other hand, the broken line shows the trend of coverage for the transcriptions. The trend of the transcriptions is similar to that for the ASR results and saturated at the 11–15th interactions. At the end of the experiments, the coverage of the transcriptions was around 75% and five points higher than ASR results. The difference in the score is due to the response selection error caused by the speech recognition results.

One possible reason for the saturated score is that the example coverage became difficult after completing the utterances of the first several interchanges. Most of the example sentences for two or three consecutive interchanges about a certain topic can be collected until around the 11–15th interaction. On the other hand, the examples of longer interchanges appear less frequently and are of a large variety. From these results, it is considered that there are few occurrences of utterances that require longer context in the non-task-oriented dialog but the dialog-based approach can collect such example sentences by iterating many interactions. As a future work, we will conduct the experiments in actual operation to investigate the degree of coverage improvement in more detail.

On the other hand, the results show that the database constructed from 25 interactions can cover 70% of the user's utterances even in the speech recognition condition. Fig. 5 shows the average coverage of transcriptions over subjects of each topic. The coverage was saturated at around the 11-15th interactions although the number of interactions at which the coverage was saturated differs slightly topic by topic. The coverage of each topic was saturated at 75% in this paper. In the following section, we compared the dialog-based databases with the handcrafted databases by subjective evaluation to investigate how the system performance improves when using the dialog-based database.

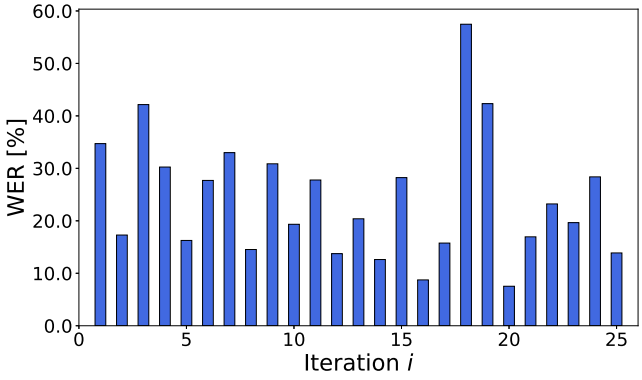

Fig. 3. WER of the speech recognition results of each interaction

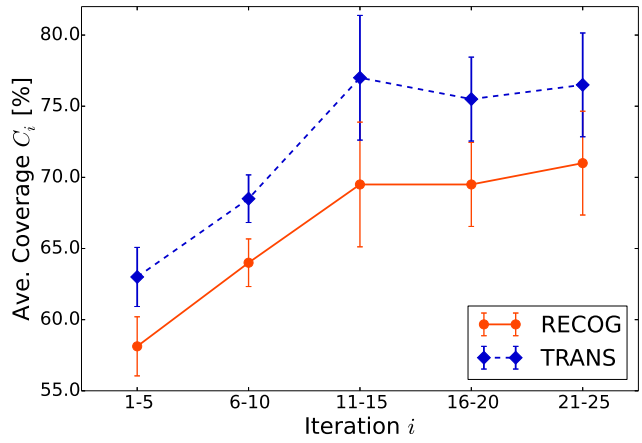

Fig. 4. Coverage of users' utterances with respect to the number of interaction

V. Comparison between dialog-based database and handcrafted database by dialog experiments

A. Construction of handcrafted database

The dialog-based databases were compared with a same-scale handcrafted database to investigate the efficacy of the example collections by the dialog experiments. In this section, we denote the database collected in Section IV as DIALOG-BASED and the handcrafted one as HANDCRAFTED in the figures and the tables. The DIALOG-

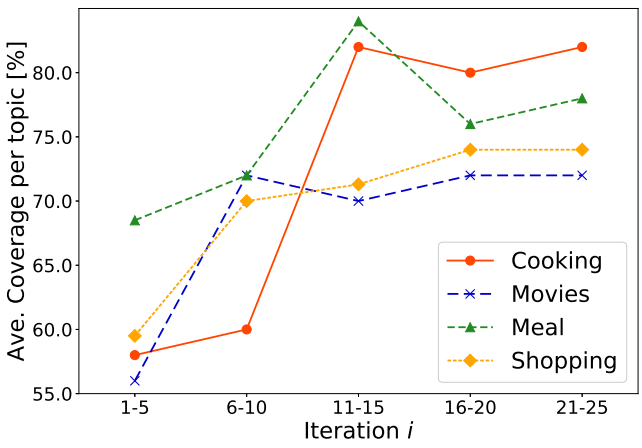

Fig. 5. Coverage of user's utterance for each topic (TRAN)[%]

TABLE V
NUMBER OF COLLECTED EXAMPLES

	Cooking	Movies	Meal	Shopping	Total
DIALOG-BASED	637	574	617	614	2442
HANDCRAFTED	625	575	625	625	2450

TABLE VI
COVERAGE OF DATABASES OF DIALOG EXPERIMENTS

	HANDCRAFTED	DIALOG-BASED
RECOG	50.0%	73.75%
TRANS	57.5%	76.00%

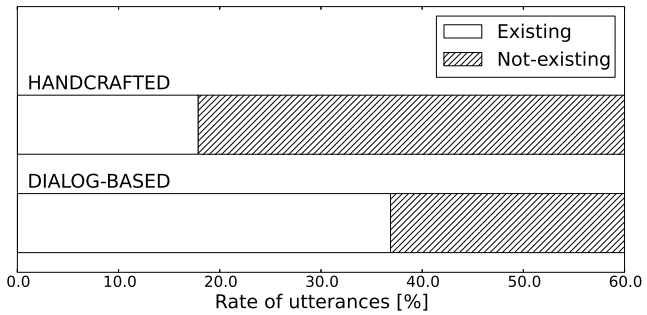

Fig. 6. Rate of user's input utterances that were exactly the same as the example

BASED of Table. V shows the number of examples collected in Section IV. As seen in the table, 2,442 examples were collected from 25 interactions.

The number and gender ratio of the creators and the number of example sentences in the handcrafted database corresponded to the dialog-based database. Therefore, the handcrafted database was created by 25 persons (15 males, 10 females). One database creator made 25 examples for three topics (cooking, meals, and shopping) and 23 examples for one topic (movies). All the database creators were presented with the initial databases, and were asked to make the examples while imagining a possible conversation under the assumption that the user talks with the agent about what happened yesterday. We collected 2,450 examples for the handcrafted database. The response sentences were developed by the same person (the first author) as the dialog-based database for consistency. Responses to examples with almost the same meanings were unified.

Here, the creators of the handcrafted database were given only the initial database, but the dialog database does not always have the advantage because the participants of Section IV talked with the system without knowing the contents of the initial database.

B. Experimental condition

Ten persons (8 males, 2 females) participated in the subjective evaluation by dialog experiments. The participants were instructed to talk with the system, regardless of whether the story continues or not, just like the experiments in Section IV. We constructed eight systems corresponding to four topics and two database creation methods of the database (i.e., dialog-based and handcrafted). Each participant talked with all the systems, and exchanged ten utterances in each interaction. They evaluated the appropriateness of each response at the end of every interaction. The order of presenting the four topics was fixed for all participants. We randomly chose two topics out of the four topics for each participant, and we presented the handcrafted-database system first for the chosen topics, and the dialog-based-database system first for the other two topics.

C. Results of objective evaluation

Firstly, we investigated the coverage of the systems. Table. VI shows the coverage of the user's utterance of each system. The table shows the average score of the dialog-based and handcrafted systems. As can be seen, the coverage of the dialog-based database was more than 70%, higher than the handcrafted one in both the speech recognition and transcription cases. These results coincide with the coverage after 25 interactions in Fig. 4. Therefore, it is shown that the database by conversation is more suited to actual use than the handcrafted database in the dialog experiment.

Then, we investigated the variation of the sentences in the databases. As mentioned above, the same response was linked to example sentences with the same meaning. Thus, the variation of the question-response pairs can be measured by counting the unique responses. The average number of unique responses and the standard deviation of the four topics were 470.5 ± 13.1 in the handcrafted database and 517.5 ± 14.9 in the dialog-based database. These results suggest that the dialog database has a large variety of examples compared to the same scale databases. Fig. 6 shows the rate of the user's input utterances that occurred in the dialog experiments that also existed in the database. As shown in the figure, the handcrafted database includes a smaller number of the same examples as the user's utterances while around 35% of the utterances could be found in the dialog database in exactly the same form. The example-based dialog system can generate a response even when the same example as the user's utterance does not exist in the database, but precise examples contribute to the generation of robust responses. This result also suggests that the example sentences created by imagining the dialog are significantly different from the actual user's utterances.

Then, we compared the number of consecutive interchanges in the conversation logs, which we defined as the number of times that the response continued to be appropriate. Table. VII shows counting of the consecutive utterances in one interaction. This example includes three consecutive interchanges and five consecutive interchanges across two not-appropriate responses. The mean number of consecutive interchanges averaged over topics is shown in Fig. 7. The error-bar is the standard error. From the unpaired t-test, we obtained a significant difference between the dialog-based database and handcrafted database ($t = -4.00, p \leq 0.01$). In the dialog with the dialog-based database system, appropriate responses continued 3.04 times on average. This result suggests that the dialog-based database includes many examples of the second or later questions compared to the handcrafted database, and the improvements of the variation of the example sentences depends on these examples. Therefore, the collection of examples by conversation is efficient, especially for the collection of interchanges of longer interaction.

TABLE VII
EXAMPLE OF COUNTING CONSECUTIVE INTERCHANGES OF ONE INTERACTION (TRANSLATION FROM JAPANESE)

No.	User	System	app.[1)]	cnt.
1.	What did you cook?	I cooked *nikujaga*.	A	3
2.	Are you good at cooking *nikujaga*?	I can't say, but I often cook it.	A	
3.	Do you make anything else?	I often cook *stir fry vegetables*.	A	
4.	Did you invite your friend?	I used *Cookpad*. Have you ever used it?	N	0
5.	Did you invite your friend?	I went for a light run after cooking.	N	0
6.	Do you often go jogging?	Yes. I'm trying to always go jogging for my health.	A	5
7.	How long did you run?	I ran about 5 km.	A	
8.	Where did you run?	The jogging path by the *Hirose river*.	A	
9.	Did you run alone?	Yes, I ran on the jogging path by the *Hirose river*.	A	
10.	Have you ever participated in a marathon?	Yes, I have participated in a half-marathon several times.	A	

[1)] A: appropriate, N: not appropriate

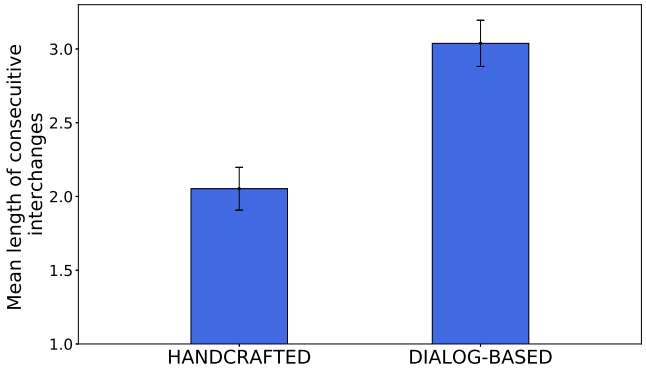

Fig. 7. Number of consecutive interchanges in dialog expresiment

Finally, the cost of database creation was compared for the two conditions. The cost was measured by the time spent to construct the database. We measured the total time to create the handcrafted database and the total dialog time for dialog-based database. The average cost to create the dialog database was 24.48 [s] and handcrafted one was 86.52 [s]. Therefore, the dialog database also has an advantage in terms of construction cost.

D. Results of subjective evaluation

For the subjective evaluation, the participants answered the following four questions based on a five-grade Likert scale from one (not at all) to five (very much).

1) **Satisfaction:** Whether the participant was satisfied with the dialog.
2) **Engagement:** Whether the participant felt that the dialog was engaged.
3) **Intelligence:** Whether the participant felt that the system was intelligent.
4) **Willingness:** Whether the participant wants to use the system again.

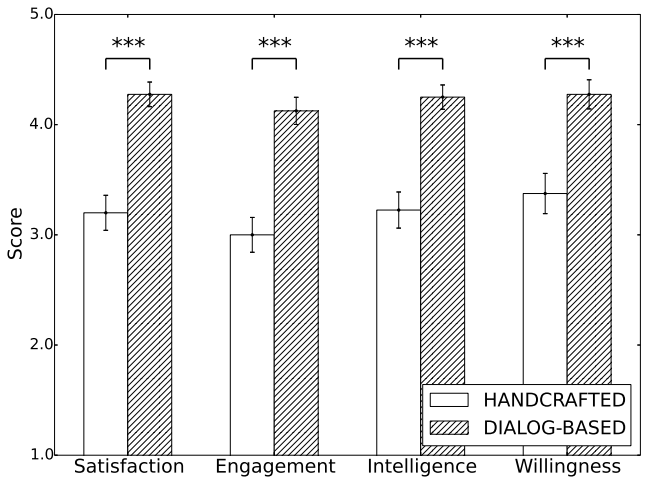

Fig. 8. Comparison of subjective evaluation scores (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

The average scores of the subjective evaluations are shown in Fig. 8. The error-bar of the figure shows the standard error. As can be seen, the dialog-based database outperformed the handcrafted database in all items. Using the unpaired t-test, we obtained a significant difference in satisfaction ($t = -5.46, p \leq 0.001$), engagement ($t = -5.54, p \leq 0.001$), intelligence ($t = -5.12, p \leq 0.001$), and willingness ($t = -3.94, p \leq 0.001$). Therefore, the dialog database also obtained a higher score than the handcrafted database in terms of the subjective evaluation. Considering the results of Section V-C, the appropriateness of the responses to consecutive utterance affected the subjective evaluation of the users.

In the non-task-oriented dialog, a long dialog is considered to be better, unlike the conventional task-oriented dialog; for example, several studies employed the dialog length as the measure for optimizing the dialog strategy [35, 36]. Examples for longer interaction are especially important for improving the user evaluation of chat-style conversation. Many of databases of conventional systems are constructed by the developer while imagining the actual dialog, but it is not easy to cover the flow of possible conversations, especially in the non-task-oriented dialog. Therefore, a framework to collect the examples by interaction is important for the future dialog system to construct an appropriate database.

E. Examples of simulated dialog of each database

Finally, we show the dialog examples in Table. VIII. "User" in the figure shows the simulated utterance sequence. "Example" and "Response" represents the example sentences and corresponding responses that are selected in the handcrafted and dialog-based databases. As shown in the table, the dialog with the dialog-based database system could select a close example to the utterance compared with a competitive system like the 1st interchange. The examined system can also respond appropriately to an utterance for which the nuance is different, but includes similar words, such as the 2nd and 3rd interchanges. In addition, the figure shows that the dialog-based database can respond appropriately when the topic changes slightly during the dialog, like the 5th interchange.

TABLE VIII
EXAMPLE OF EXAMPLE-RESPONSE PAIRS FOR SIMULATED UTTERANCE SEQUENCE (COOKING)

			HANDCRAFTED		DIALOG-BASED	
No.	Lang.	User	Example	Response	Example	Response
1.	Jp	Nikujaga wa oishikatta desu ka?	Nikujaga wa tokui desu ka?	Nikujaga wa tokuiryôri to wa ie mase n ga, tsukuru koto wa ôi desu.	Nikujaga wa oishikatta desu ka?	Hai. oishiku deki mashi ta.
	En	How was *nikujaga*?	Are you good at cooking *nikujaga*?	I can't say, but I often cook it.	How was *nikujaga*?	Yes. It was delicious.
2.	Jp	Ryôri wa tokui na n de su ka?	Tokui na ryôri wa nan desu ka?	Hoikôrô ga tokui desu.	Ryôri wa tokui na n desu ka?	Udemae wa futsû da to omoi masu. demo, tsukuru koto wa daisuki desu.
	En	Are you good at cooking?	What is your specialty?	*Twice cooked pork*.	Are you good at cooking?	It's OK, but I love cooking.
3.	Jp	Ryôri igai ni suki na koto wa ari masu ka?	Nani ka hoka ni tokui na ryôri wa ari masu ka?	Omuraisu mo tokui ryôuri desu yo.	Ryôri igai ni suki na koto wa ari masu ka?	Undô ya eiga kanshô ga suki desu.
	En	Do you have any favorite things other than cooking?	Do you have any other specialty?	*A rice omelet*.	Do you have any favorite things other than cooking?	I like to do exercise or watch movies.
4.	Jp	Ryôri wo tsukutta ato wa nani wo shi mashi ta ka?	Ryôri wo tsukutta ato wa nani wo shi mashi ta ka	Karuku ranningu wo shi mashi ta.	Ryôri shi ta ato wa nani ka shi mashi ta ka?	Ryôri wo shi ta ato wa, karuku ranningu wo shi mashi ta.
	En	What did you do after cooking?	What did you do after cooking?	I went for a light run.	What did you do after cooking?	I went for a light run after cooking.
5.	Jp	Doko wo hashiri mashi ta ka?	Doko no reshipi wo mi mashi ta ka?	Kukkupaddo wo tsukai mashi ta. anata mo tsukatta koto wa ari masu ka?	Doko wo hashiri mashi ta ka?	Hirosegawa zoi no ranningu kôsu wo hashitte i mashi ta.
	En	Where did you run?	Which recipe did you use?	I used *Cookpad*. Have you used it?	Where did you run?	Along the *Hirose river*.

VI. CONCLUSION

In this paper, we examined an example collection method by conversation for a non-task-oriented example-based dialog system, and we showed the efficacy of the method by several analyses. In the examined approach, the database was updated at the end of every interaction. The trend of the appropriateness of the responses showed that a database constructed from 25 interactions can cover close to 70% of the user's utterance even in the speech recognition condition. In addition, analyses of the collected examples suggested that the dialog-based database responds appropriately to consecutive utterances compared with the handcrafted database. The examined approach outperformed even in the subjective evaluation. These results show that the dialog-based approach is superior to the handcrafted approach not only in terms of coverage but also in the subjective evaluation.

One of the remaining issues is that part of the processing of this study is conducted manually. As a future work, we will examine methods to decide the user's utterances for assigning the database and creating the responses automatically.

REFERENCES

[1] T.W. Bickmore and R.W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions on Computer-Human Interaction*, vol. 12, no. 2, pp. 293–327, 2005.

[2] T. Meguro, R. Higashinaka, Y. Minami and K. Dohsaka, "Controlling listening-oriented dialogue using partially observable Markov decision processes," in *Proc. of the 23rd International Conference on Computational Linguistics*, pp. 761–769, 2010.

[3] Z. Yu, L. Nicolich-Henkin, A. W. Black and A.I. Rudnicky, "A wizard-of-Oz study on a non-task-oriented dialog systems that reacts to user engagement," in *Proc. of the 17th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 55–63, 2016.

[4] H. Aust, M. Oerder, F. Seide and V. Steinbiss, "The Philips automatic train timetable information system," *Speech Communication*, vol. 17, no. 3–4, pp. 249–262, 1995.

[5] V. Zue, S. Seneff, J.R. Glass, J. Polifroni, C. Pao, T.J. Hazen and L. Hetherington, "JUPITER: a telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 85–96, 2000.

[6] T. Bickmore and J. Cassell, "Relational agents: a model and implementation of building user trust," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 396–403, 2001.

[7] R. Nisimura, A. Lee, H. Saruwatari and K. Shikano, "Public speech-oriented guidance system with adult and child discrimination capability," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I, pp. 433–436, 2004.

[8] D. Traum, K. Georgila, R. Artstein and A. Leuski, "Evaluating spoken dialogue processing for time-offset interaction," in *Proc. of the 16th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 199–208, 2015.

[9] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communication of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[10] R. Wallace, *Be your own botmaster*, A.L.I.C.E A.I. Foundation, 2003.

[11] C. Lee, S. Jung, S. Kim, and G. G. Lee, "Example-based dialog modeling for practical multi-domain dialog system," *Speech Communication*, vol. 51, no. 5, pp. 466–484, 2009.

[12] R. Dickerson, K. Johnsen, A. Raij, B. Lok, J. Hernandez, A. Stevens, and DS Lind, "Evaluating a script-based approach for simulating patient-doctor interaction," in *Proc. the International Conference of Human-Computer Interface Advances for Modeling and Simulation*, pp. 79–84, 2005.

[13] P. Kenny, T.D. Parsons, J. Gratch, and A.A. Rizzo, "Evaluation of Justina: A Virtual Patient with PTSD," *Intelligent Virtual Agents*, pp. 394–408, 2008.

[14] D.G. Bobrow, R.M. Kaplan, M. Kay, D.A. Norman, H. Thompson and T. Winograd, "GUS, a frame-driven dialog system," *Artificial Intelligence*, vol. 8, no. 2, pp. 155–173, 1977.

[15] H. Sugiyama, T. Meguro, R. Higashinaka and Y. Minami, "Large-scale collection and analysis of personal question-answer pairs for conversational agents," *International Conference on Intelligent Virtual Agents*, pp. 420–433, 2014.

[16] A. Ritter, C. Cherry and B. Dolan, "Unsupervised modeling of Twitter conversations," in *Proc. of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 172–180, 2010.

[17] F. Bessho, T. Harada and Y. Kuniyoshi, "Dialog system using real-time crowdsourcing and twitter large-scale corpus," in *Proc. of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 227–231, 2012.

[18] H. Chung, Y.I. Song, K.S. Han, D.S. Yoon, J.Y. Lee, H.C. Rim, and S.H. Kim, "A practical QA system in restricted domains," in *Proc. the Workshop on Question Answering in Restricted Domains*, pp. 39–45, 2004.
[19] B. Katz, "Annotating the world wide web using natural language," in *Proc. of the RIAO Conference on Computer-Assisted Information Searching on Internet*, pp. 136–155, 1997.
[20] H. Murao, N. Kawaguchi, S. Matsubara, Y. Yamaguchi, and Y. Inagaki, Example-based spoken dialogue system using WOZ system Log, in *Proc. of the 4th SIGDIAL Workshop on Discourse and Dialogue*, pp. 140–148, 2003
[21] L. Nio, S. Sakti, G. Neubig, T. Toda, M. Adriani, and S. Nakamura, "Developing non-goal dialog system based on examples of drama television," *Natural Interaction with Robots, Knowbots and Smartphone*, pp. 355–361, 2013.
[22] R.E. Banchs and H. Li, "IRIS: a chat-oriented dialogue system based on the vector space model," in *Proc. of the ACL 2012 System Demonstrations*, pp. 37–42, 2012.
[23] R. Yan, Y. Song and H. Wu, "Learning to respond with deep neural networks for retrieval-based human-computer conversation system," in *Proc. of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 55–64, 2016.
[24] A. Ritter, C. Cherry and B.D. Dolan, "Data-driven response generation in social media" in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 583–593, 2011.
[25] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015
[26] B. Rossen and B. Lok, "A crowdsourcing method to develop virtual human conversational agents," *International Journal of Human-Computer Studies*, vol. 70, no. 4, pp. 301–319, 2012.
[27] A. Lee, K. Oura and K. Tokuda, "MMDAgent—A fully open-source toolkit for voice interaction systems," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8382–8385, 2013.
[28] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proc. of APSIPA ASC*, pp. 131–137, 2009.
[29] Open JTalk, http://open-jtalk.sourceforge.net/.
[30] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan*, vol. 20, no. 3, pp. 199–206, 1999.
[31] K. Maekawa, H. Koiso, S. Furui and H. Isahara, "Spontaneous Speech Corpus of Japanese," in *Proc. of LREC*, pp. 947–952, 2000.
[32] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.
[33] H. Kawahara, I Masuda-Katsuse and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.
[34] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. of INTERSPEECH*, pp. 1393–1396, 2004.
[35] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, "Deep Reinforcement Learning for Dialogue Generation," in *arXiv preprint arXiv:1606.01541*, 2016
[36] Z. Yu, Z. Xu, A.W. Black, and A.I. Rudnicky, "Strategy and Policy Learning for Non-Task-Oriented Conversational Systems," in *Proc. of the 17th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 404–412, 2016.