User Impressions of Questions to Acquire Lexical Knowledge

Kazunori Komatani
Osaka University
8-1 Mihogaoka, Ibaraki
Osaka 567-0047, Japan
komatani@sanken.osaka-u.ac.jp

Mikio Nakano
Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako
Saitama 351-0188, Japan
nakano@jp.honda-ri.com

Abstract

For the acquisition of knowledge through dialogues, it is crucial for systems to ask questions that do not diminish the user's willingness to talk, i.e., that do not degrade the user's impression. This paper reports the results of our analysis on how user impression changes depending on the types of questions to acquire lexical knowledge, that is, explicit and implicit questions, and the correctness of the content of the questions. We also analyzed how sequences of the same type of questions affect user impression. User impression scores were collected from 104 participants recruited via crowdsourcing and then regression analysis was conducted. The results demonstrate that implicit questions give a good impression when their content is correct, but a bad impression otherwise. We also found that consecutive explicit questions are more annoying than implicit ones when the content of the questions is correct. Our findings reveal helpful insights for creating a strategy to avoid user impression deterioration during knowledge acquisition.

1 Introduction

Structured knowledge bases are not only crucial for providing various services such as information search and recommendations but also effective for non-task-oriented dialogue systems to avoid generic or dull responses (Xing et al., 2017; Young et al., 2018; Zhou et al., 2018; Liu et al., 2019). However, it is impractical to presuppose a perfect knowledge base (West et al., 2014) in an early stage of system development.

Therefore, being able to acquire knowledge from users and thereby enhance knowledge bases through dialogues is one of the most important abilities that dialogue systems should possess. Although knowledge acquisition can be done by asking people to input information on GUIs or spreadsheets, knowledge acquisition through dialogues has an advantage in that people can enjoy conversations with the system, especially when the system can engage in non-task-oriented dialogues (Kobori et al., 2016).

Figure 1: An example of implicit confirmation.

One of the targets of knowledge acquisition through dialogues is knowledge about unknown terms and unknown relations between terms by asking the appropriate questions. This would enable the systems to keep learning even when unknown terms appear during dialogues (Meena et al., 2012; Sun et al., 2015).

To enable non-task-oriented systems to acquire a variety of knowledge, the dialogue needs to continue, but this can be a difficult task, as revealed in the Amazon Alexa Prize challenges (Fang et al., 2018; Chen et al., 2018). Users tend to stop interacting with a dialogue system if it repeatedly asks annoying questions, as they do not wish to use the system like an "oracle" who must repeatedly tell it whether a target is correct or wrong (Amershi et al., 2014). Therefore, asking questions for acquiring knowledge should be designed so that they do not irritate the user too much.

For acquiring domain knowledge without asking abrupt questions, the process of implicit confirmation was proposed for non-task-oriented dia-

, pages 138 - 147

logue systems (Ono et al., 2016, 2017). Figure 1 shows an example of this process. First, when an unknown term appears in a user utterance, the system estimates its attribute (Otsuka et al., 2013) (Step 1). Second, the system asks an implicit question[1] about the estimated result, instead of asking an explicit question (Step 2). The implicit question is not a superficially interrogative sentence, but it functions as a question by interpreting it along with the subsequent user utterance. Third, the system determines whether or not the estimated result included in the implicit question was correct by also taking the subsequent user response into consideration, and then it adds the estimated result to the system knowledge if it is correct (Step 3). Although these studies assume that implicit questions are less irritating than explicit questions, this has not been empirically verified. Moreover, since the estimated results used in the questions are not always correct, any effect on user impression when the results in the questions are wrong should be considered.

We therefore investigate how system questions for acquiring knowledge affect user impression, including the user's irritation by asking the extent to which the system utterances were annoying. Here, two research questions, RQ1 and RQ2, are addressed. RQ1 is how the system's question types affect user impression. The questions consist of five types comprising both explicit and implicit questions, and the correctness of the content of the questions. RQ2 is whether or not consecutive explicit questions for acquiring knowledge are felt as more annoying than consecutive implicit ones. A strategy based upon the results will be discussed in Section 5.

We gathered user impression data after users engaged in a session consisting of several interactions with the system and then analyzed the impression in relation to the question types used in the session. The most naive approach to obtain user impressions is to ask after every system turn, but this would be very annoying and disturb the dialogue flow. Instead, we estimated the effect of each question type in the session by means of a regression model. This model also enables us to analyze user impression when the same question type is repeated.

[1]This system utterance was called an implicit *confirmation request* in (Ono et al., 2016, 2017), but in this paper we call it an implicit *question* to clarify their difference in purpose, which will be explained in Section 2.2.

2 Related Work

2.1 Knowledge acquisition in dialogue systems

It has been of great interest that computers continue to learn knowledge autonomously. A famous example is the Never-Ending Language Learner (NELL) (Carlson et al., 2010; Mitchell et al., 2015), which continuously extracts information from the Web. Several methods have been developed for machine learning tasks (such as information extraction) to continuously improve the performance of classifiers in a semi-supervised manner, which is known as life-long learning (Chen and Liu, 2018). We aim to develop systems that can perform such knowledge acquisition through dialogues.

Several studies have investigated how dialogue systems acquire knowledge. Otsuka et al. (2013) proposed a method to estimate the cuisine of an unknown restaurant name from its character sequence and to accordingly change question forms to acquire knowledge. Pappu and Rudnicky (2014) designed strategies for asking users questions in a goal-oriented dialogue system and analyzed the acquired knowledge through a user study. Hixon et al. (2015) proposed a method for asking questions to obtain relations between concepts in a question-answering system. Weston (2016) designed ten tasks and demonstrated that supervision given as feedback from simulated interlocutors enables an end-to-end memory network to predict the next utterances better; Li et al. (2017) implemented Weston's method with reinforcement learning and showed that the system performance improved by asking questions. Mazumder et al. (2019) proposed a system that asks questions about a triple by using knowledge graph completion where a triple (s, r, t) denotes a source entity, a relation, and a target entity, respectively, and lacks either a source s or target t. In these problem settings, it is important to consider how users feel about the system's questions in order to continue dialogues to acquire a variety of knowledge. As mentioned in Section 1, Ono et al. (2017) proposed implicit questions to avoid decreasing the user's willingness to talk, but its effect has not been verified through a user study.

2.2 Implicit questions

Implicit questions for non-task-oriented dialogues (Ono et al., 2017) differ from implicit confirma-

	Correct **C**	Wrong **W**
Explicit **E**	**EC** "Is puttanesca Italian?"	**EW** "Is puttanesca Japanese?"
Implicit **I**	**IC** "Italian is perfect for a date."	**IW** "Japanese foods are healthy."
Whq	**Whq** "What is puttanesca?"	

Table 1: Examples of five types of system questions for *puttanesca* whose correct cuisine is *Italian*.
E and I denote explicit and implicit questions. C and W denote whether the content is correct or wrong. Whq denotes Wh-questions.

tion requests for task-oriented dialogues from the viewpoint of purpose. Implicit confirmation is a well-known technique for task-oriented spoken dialogue systems as a way of handling errors (Bohus and Rudnicky, 2005; Skantze, 2005). A number of studies have focused on changing the form of confirmation requests, including explicit and implicit ones (Bouwman et al., 1999; Komatani and Kawahara, 2000). Consider an example in a flight reservation task where the system tries to determine the destination (going to Boise). The system can ask something like "Are you going to Boise?" as an explicit confirmation request, and it can also continue the dialogue by asking its next question, e.g., "To get to Boise, where will you depart from?", as an implicit confirmation request. Prior research in task-oriented dialogues has shown that an implicit confirmation request can reduce the number of turns when the content is correct and that it is difficult to correct the system's misunderstanding when the content is incorrect (Sturm et al., 1999).

The advantage of implicit questions in non-task-oriented dialogues is not the reduction in the number of turns, which is well-known in task-oriented dialogues, but rather that they do not disturb the dialogue flow, which hopefully will decrease the likelihood of the user becoming irritated and stopping the dialogue. User impression, particularly how annoying a question type is, should be investigated in order to enable non-task-oriented systems to continue dialogues, especially when they are utilized by real users. In this paper, we address this issue from the viewpoint of user impression through a user study.

2.3 User impression of dialogues

Several studies have tried to predict user impression of dialogues. Walker et al. (1997) proposed a framework to predict user satisfaction by means of a regression model using various objective factors during task-oriented dialogues. Higashinaka et al. (2010) developed a method to model user satisfaction transitions using a hidden Markov model even when only user impression scores for entire dialogues were given. Ultes and Minker (2014) and Ultes (2019) improved the prediction accuracy of the interaction quality with various machine learning methods. In contrast, the aim of this paper is *not* to predict user impressions, but rather to analyze the effects of question types on them in dialogues by means of a regression model inspired by (Walker et al., 1997).

3 User Study Design

We assume a system that obtains an attribute value for an unknown term. That is, when an unknown term appears in a dialogue, we try to make the system acquire its attribute from the user through the dialogue. A pair consisting of the term and its attribute can then be stored as new system knowledge.

More specifically, we assume the pair of an unknown food name and its cuisine. First, the cuisine of a food name is estimated from its character sequence (Otsuka et al., 2013), and next, the estimated cuisine is verified by asking either form of question. We focus here on the types of questions for verifying the estimated cuisine.

3.1 Five question types for knowledge acquisition

Table 1 lists the five question types along with examples. The examples correspond to a case where the unknown term is *puttanesca*, its estimated correct cuisine is *Italian*, and its estimated wrong cuisine is *Japanese*.

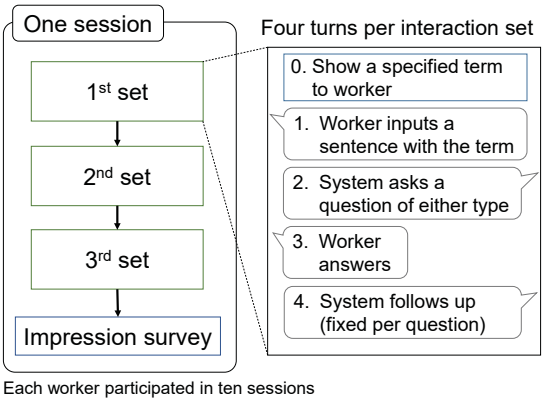

Figure 2: Flow of data collection.

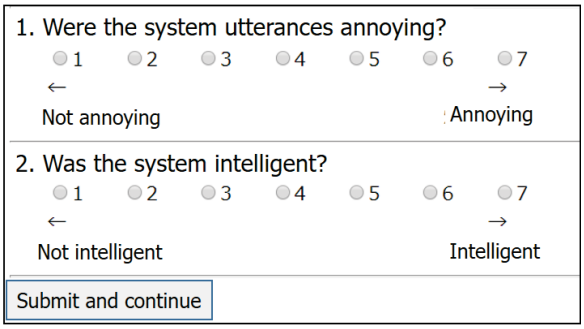

Figure 3: Screenshot of impression survey.

The question types have two components: its question form and the correctness of its content. The first one can be explicit ('E'), implicit ('I'), or a Wh-question ("Whq"). An explicit question explicitly asks whether its content is correct or not through a Yes/No question (e.g., "Is puttanesca Italian?"). An implicit question continues the dialogue with a system utterance containing the estimated cuisine name (e.g., "Italian is perfect for a date.") and then implicitly determines whether the cuisine is correct or not by also considering the subsequent user utterance (Ono et al., 2017). A Wh-question simply asks without using an estimated cuisine (e.g., "What is puttanesca?").

The other component is whether the estimated cuisine is correct or not. We utilize it to investigate any effects on user impression caused by correct or wrong content, which is derived from the automatic estimation about the unknown food name (Otsuka et al., 2013), before the system asks a question. This is applied only to the explicit and implicit questions, as Wh-questions have no concrete content. Thus, C and W, which respectively denote correct and wrong content, are added to E and I explained above, except for Whq. For simplicity, we only consider explicit questions with one choice and do not consider those with multiple choices (Komatani et al., 2016).

3.2 Data collection

We investigated user impression of dialogues including questions of the five types via crowdsourcing.[2] Crowdworkers were Japanese speakers and thus all the dialogues were in Japanese. We explained that they would talk with an "AI chatbot" and asked them to talk as if they were meeting for the first time.

The workers gave their impression scores once per session. The flow is depicted in Figure 2. One session consists of three sets of interactions, followed by an impression survey.

Each interaction set consists of four turns: two system turns and two user turns. Before the first turn, a term is displayed as an instruction, e.g., "Please input your thought as if you ate *puttanesca* recently". The four turns flow as follows.

Turn 1: A worker inputs a sentence containing the term specified in the instruction. The term is prepared before the experiment.

Turn 2: The system asks a question about the term as one of the five question types. The question type is randomly selected from the five. Wrong cuisine estimation results and expressions of implicit questions are manually prepared before the experiment.

Turn 3: The worker inputs a response to the system question. There is no restriction on the response.

Turn 4: The system's follow-up response is displayed. It depends on the question type[3] used in Turn 2. For example, it is "Sorry, I probably misunderstood." for type IW (implicit, wrong).

One interaction set ends after the four turns have finished, and then the next specified term is displayed for the next interaction set.

After engaging in the interaction sets three times, the workers fill in a questionnaire (Figure 3)

[2]We used the platform of CrowdWorks, Inc. https://crowdworks.co.jp/

[3]The type selection is random. It is not based on automatic estimation, e.g., determining whether the content is correct or wrong, in this data collection.

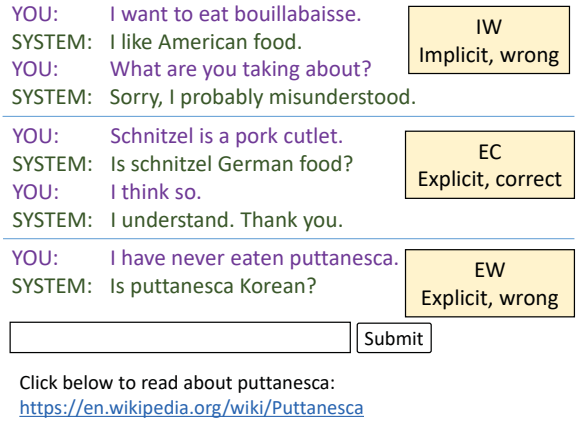

Figure 4: A screenshot during a third set. The boxes on the right with the question types are only for explanation and were not displayed.

about their impression scores for the session. The questionnaire features 7-point Likert scales for "Were the system utterances annoying?" and "Was the system intelligent?".[4] Hereafter, these impression scores are denoted as *annoying* and *intelligent*, respectively.

Each worker was asked to engage in ten sessions. The number of specified terms, which are regarded as unknown terms, was 30, that is, three per session.

Figure 4 shows an example screenshot (translated from Japanese). The lines starting with "YOU" and "SYSTEM" denote a worker's and the system's utterances, respectively. The initial part of each interaction set, in which the specified term was shown to workers, is not displayed in the figure, as it disappears when workers input their first sentence. If a worker did not know the term, he or she could check Wikipedia via a link at the bottom of the screen. This was to prevent dialogues in which workers were unaware of the term's meaning. The dialogues are not very natural, but we used them as the first step for this kind of study, since currently there is no system that can acquire knowledge many times in a natural way.

In total, we obtained 1,183 sessions by 104 workers after removing unusable data (e.g., that of workers who did not finish all ten sessions) from the original 1,319 sessions by 120 workers.[5]

[4]These questionnaire items are unvalidated; they are not captured using redundancy (i.e., different ways of asking the same content) in order to minimize misinterpretations, as argued in (Davis, 1989). We used simple items because they were easiest to explain to the crowdworkers.

[5]Due to a system error, some workers engaged in more than ten sessions.

That is, we obtained 1,183 *annoying* and *intelligent* impression scores corresponding to every session, each of which contains three system question types to be analyzed. There was little agreement among the workers because the impression scores are subjective; some workers gave higher scores overall and others did the opposite. However, there is a certain tendency within each worker's impression scores for different question types.

4 Analysis with Linear Regression

We analyzed the effect of each question type by using the coefficients of a linear regression model that predicts the collected impression scores. First, we describe the basic regression model and its refinement to make the multiple correlation coefficients (R) higher. After that, we discuss the effect of each question type on user impression and analyze results when the same question types were repeated.

4.1 Linear regression model

A linear regression model was used to predict user impression scores (*annoying* or *intelligent*) from the number of question types used in each session. The basic regression model for the score of the i-th session is given as

$$\text{score}_i = w_0 + \sum_{c \in \{EC,EW,IC,IW,Whq\}} w_c n_i(c), \quad (1)$$

where $n_i()$ denotes the number of each question type c used in the session. The value was 0, 1, 2, or 3 in the basic model.

We applied two refinements to improve the multiple correlation coefficients. First, we normalized impression scores to make their mean 0 and variance 1 per worker. This is effective because each worker gave impression scores in a different range; that is, some gave higher scores on average on the 7-point scale, while others gave lower. As we wanted to know the effect of each question type that had been randomly selected, we used the relative scores given by each worker.

Second, we considered the temporal position of the questions out of the three interaction sets in a session. That is, we used 15 independent variables: the five question types having the three positions each (representing the first, second, and third interaction sets in one session). The refined

	intelligent	*annoying*
Basic regression model	0.368	0.207
+Normalized per worker	0.493	0.308
+Considering positions	0.540	0.354

Table 2: Multiple correlation coefficients (R) of the models.

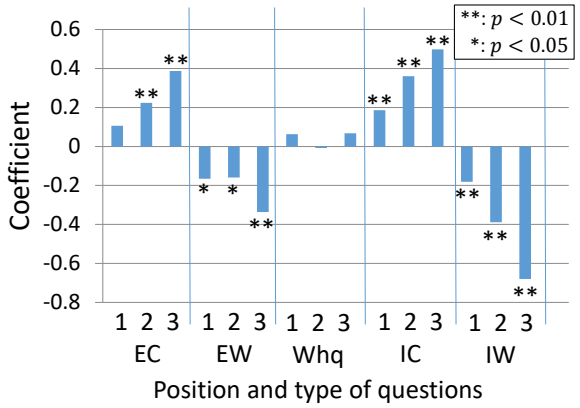

Figure 5: Coefficients of the regression model for *intelligent* when types and positions were considered. The symbols ** and * denote the coefficient is not zero with statistical significance at $p < 0.01$ and $p < 0.05$, respectively.

regression model is given as

$$\text{score}_i = w_0 + \sum_d w_d n_i(d), \qquad (2)$$

where $d \in \{EC, EW, IC, IW, Whq\} \times \{1, 2, 3\}$, and $n_i()$ denotes the number of each question type with the position d. It is thus binary in this refined model.

The multiple correlation coefficients for the two impression scores are listed in Table 2. The coefficients became higher by the normalization, and became even higher by considering the temporal positions. Thus, in the following analysis, we use the model with these 15 coefficients considering the positions after the normalization per worker.

The table also shows that the *intelligent* scores had a better fit to the collected data. Since the two impression scores had almost the reverse tendency, either will be used in the following sections for brevity.

4.2 Analysis of obtained coefficients

RQ1 is addressed here: "how the system's question types affect user impression". Figure 5 shows the values of the 15 coefficients obtained for *intelligent*, which fitted the data better. We also tested the statistical significance of individual regression coefficients that verifies whether or not the coefficient is zero; these results are shown as well. Larger positive values indicate that the question type in that position tends to give a better impression to workers, that is, they felt the system was more intelligent. Larger negative values indicate the opposite.

	EC	EW	Whq	IC	IW
intelligent	0.24	−0.22	0.04	0.35	−0.42
annoying	−0.13	0.08	−0.02	−0.21	0.28

Table 3: Averages over the three positions of the coefficients in the regression models.

The averages over the three positions for *intelligent* are summarized in Table 3, along with those for *annoying*. The coefficients of the types are ordered as

$$\text{IC} > \text{EC} > \text{Whq} > \text{EW} > \text{IW}$$

for *intelligent*, and

$$\text{IC} < \text{EC} < \text{Whq} < \text{EW} < \text{IW}$$

for *annoying*. The two impression scores showed the reverse order.

Details follow using the case of *intelligent*. The coefficients of IC and EC, both of which had correct content, were positive, and those of EW and IW, both of which had wrong content, were negative. This result corresponds to our intuition that workers feel the system is not intelligent when it asks questions with the wrong content. The coefficient of Whq was in-between, as it had no concrete content.

Next, we focus on the relationship between the explicit and implicit questions. When they had correct content, the coefficients of the implicit questions (IC) were larger than those of the explicit questions (EC). This result indicates that the implicit questions give a better impression than the explicit ones. This is because the workers felt the system knew rare and difficult terms; the impression scores were higher when the target food names seemed more uncommon. In contrast, when they had wrong content, the coefficients of the explicit questions (EW) were less negative than those of the implicit questions (IW). In other words, if the estimated cuisine was wrong, the explicit questions caused less damage to user impression than the implicit ones. This is probably because the workers felt the system ignored their previous utterances and selfishly started a new topic

when an implicit question was asked with a wrong cuisine.

Figure 5 also shows the tendency among three temporal positions of each question type. In the cases of both negative and positive coefficients, they were the largest at the third positions for all five types. This suggests that the question type just before the impression survey might have the largest effect on the impression scores.

4.3 Impression when the same question type is repeated

This section addresses RQ2: "whether or not consecutive explicit questions are considered more annoying than implicit ones". Here, the impression scores for *annoying* are used, as the purpose of RQ2 is to investigate whether the consecutive questions are annoying or not.

We compare the following two impression scores for the case where the same question type is repeated.

- Actual scores when same question type was repeated three times
- Predicted scores by regression model

By comparing the two scores, we can analyze the difference between impression when the same question type was actually repeated and that when the question type was used with various contexts.

Specifically, the former scores were calculated by averaging the scores of the sessions where the same question types were actually repeated as a result of random selection. Such cases occurred 10.4 times on average per question type in the collected data. On the other hand, the latter scores were calculated with the model of Eq. (2) for the cases when a question type was used three times. Its coefficients were obtained using data where each question type was randomly selected, that is, without considering whether the same question types were repeated or not. We can thus regard them as averages over the cases when the five question types appeared in various contexts.

Figure 6 shows the results and Table 4 lists their concrete values. For all question types, the impression scores for the actual cases were larger, i.e., more annoying, than those for the predicted cases. Furthermore, the differences in the scores for types with wrong content (EW and IW) were larger than those with correct content (EC and IC), as shown in the "Difference" column in Table 4.

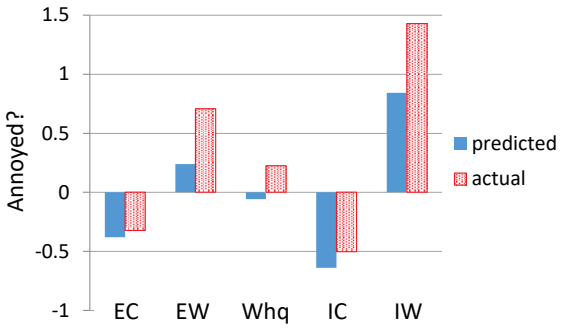

Figure 6: Scores for *annoying* (predicted vs. actual).

	Predicted	Actual	Difference
EC	−0.380	−0.324	+0.056
EW	0.239	0.707	+0.468
Whq	−0.058	0.224	+0.282
IC	−0.639	−0.502	+0.137
IW	0.842	1.429	+0.587

Table 4: Predicted and actual impression scores for *annoying* when the same question types are repeated three times.

This corresponds to our intuition that consecutive questions with wrong content would be more annoying than consecutive ones with correct content.

The results also show that the scores of IC were more negative, i.e., less annoying, than those of EC in both the predicted and actual cases. Although the degrees of "not annoying" diminished in the actual cases, the order between them did not change. This means that, when the content is correct, consecutive explicit questions (EC) are more annoying than consecutive implicit ones (IC). This result answers RQ2; it is true when the content is correct.

One reason for these results is that all explicit questions had almost the same simple form (e.g., "Is puttanesca Italian?"), and this made users feel as if the system repeated such questions without thinking. In contrast, implicit questions disturbed the dialogue flow less and thus were less annoying, as they contained the content of the previous user utterance when the cuisine in it was correct.

Example dialogues of sessions when explicit and implicit questions with correct content (EC and IC) were asked three times are respectively shown in Figures 7 and 8.

5 Conclusion

We investigated how five types of question affect user impression and clarified the order of the ques-

User	The *panna cotta* was very sweet and good.
System	Is panna cotta Italian?
User	Yes. It's Italian.
System	I understand. Thank you.
User	The *pad thai* was refreshing and delicious.
System	Is pad thai a Thai food?
User	That's right. It's a Thai food.
System	I understand. Thank you.
User	I ate a *bagna cauda*. It was similar to shabu-shabu.
System	Is bagna cauda Italian?
User	Yes. It is an Italian pot cuisine.
System	I understand. Thank you.

Figure 7: Example session of consecutive explicit questions with correct content (EC).

User	*Gelato* looks gorgeous.
System	Ice confections are perfect for eating on a hot day.
User	That's right. It's refreshing.
System	I think so too.
User	I ate *paella* for dinner.
System	Many Japanese people like Spanish food.
User	Japanese people like such taste.
System	I think so too.
User	*Ushiojiru* is a nice, tasty soup.
System	Many Japanese dishes are lightly seasoned.
User	That's right. They're so good.
System	I think so too.

Figure 8: Example session of consecutive implicit questions with correct content (IC).

tion types in terms of two impression scores: *intelligent* and *annoying*. We also found that repeating the same question type annoys the user and degrades user impression even when the content is correct. The implicit questions tend to give a better user impression than the explicit ones as long as the content is correct. On the other hand, the impression by the implicit questions with wrong content is worse.

These results justify an adaptive strategy using a confidence measure of the cuisine estimation in order to lower the risk of user impression deterioration. Specifically, it is reasonable that the system should select an implicit question if the confidence is high, and select an explicit one (or a Wh-question) if the confidence score is low. This is consistent with results discussed for task-oriented dialogues using confidence measures of automatic speech recognition, such as (Sturm et al., 1999).

A new finding here, based on the results of our analysis in Section 4.3, is that the designer of the dialogue system also needs to avoid repeating the same type of questions in non-task-oriented dialogues. The system should have multiple choices of question types in order to prevent users from becoming irritated. That is, it is necessary to change question types appropriately by considering not only the confidence of the estimation but also the history of the dialogue. This will help the dialogue to continue with less degradation of the user's impression and enable the system to acquire knowledge through dialogues.

Several issues remain as future work. Our experiment was limited in terms of the number of turns and the domain where it was tested. The results need to be verified with non-task-oriented systems that can engage in longer dialogues in various domains. We are planning to implement a non-task-oriented dialogue system that has the function to acquire knowledge. The subdialogue shown in this paper can be embedded within a longer dialogue. The implicit confirmation can be implemented by preparing the expressions of implicit questions for each category (cuisine type, in this paper) to be estimated. A further user study will be conducted with the implemented system. Another issue is that answers from users may be different; e.g., some users may say that "mapo doufu" is Sichuan, but others may say it is Chinese. This is caused by the different concept granularity of individual users, which appears in the answers. A knowledge graph that can have different nodes representing the both concepts may be a possible solution for this issue. Incorporating the utility of each question type for acquiring knowledge (Komatani et al., 2016) would be another interesting extension of the strategy.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JP16H02869 and JP19H04171.

References

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4).

Dan Bohus and Alexander Rudnicky. 2005. Error handling in the RavenClaw dialog management architecture. In *Proc. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 225–232.

Gies Bouwman, Janienke Sturm, and Lou Boves. 1999. Incorporating confidence measures in the Dutch

train timetable information system developed in the ARISE project. In *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proc. Conference on Artificial Intelligence (AAAI)*.

Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Runxiang Cheng, Ashwin Bhandare, and Zhou Yu. 2018. Gunrock: Building a human-like social bot by leveraging large scale real user data. In *2nd Proceedings of Alexa Prize (Alexa Prize 2018)*.

Zhiyuan Chen and Bing Liu. 2018. *Lifelong Machine Learning, Second Edition*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.

Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A. Smith, and Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)*, pages 96–100.

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Modeling user satisfaction transitions in dialogues from overall ratings. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, page 18–27.

Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning knowledge graphs for question answering through conversational dialog. In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)*, pages 851–861.

Takahiro Kobori, Mikio Nakano, and Tomoaki Nakamura. 2016. Small talk improves user impressions of interview dialogue systems. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 370–380.

Kazunori Komatani and Tatsuya Kawahara. 2000. Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output. In *Proc. International Conference on Computational Linguistics (COLING)*, pages 467–473.

Kazunori Komatani, Tsugumi Otsuka, Satoshi Sato, and Mikio Nakano. 2016. Question selection based on expected utility to acquire information through dialogue. In *Proc. International Workshop on Spoken Dialogue System Technology (IWSDS)*, pages 27–38.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Learning through dialogue interactions by asking questions. In *Proc. International Conference on Learning Representations (ICLR)*.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proc. Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1782–1792.

Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. 2019. Lifelong and interactive learning of factual knowledge in dialogues. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 21–31.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2012. A data-driven approach to understanding spoken route directions in human-robot dialogue. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 226–229.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *Proc. Conference on Artificial Intelligence (AAAI)*.

Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. 2016. Toward lexical acquisition during dialogues through implicit confirmation for closed-domain chatbots. In *Proc. of Second Workshop on Chatbots and Conversational Agent Technologies (WOCHAT)*.

Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. 2017. Lexical acquisition through implicit confirmations over multiple dialogues. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 50–59.

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

Aasish Pappu and Alexander Rudnicky. 2014. Knowledge acquisition strategies for goal-oriented dialog systems. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 194–198.

Gabriel Skantze. 2005. Galatea: a discourse modeller supporting concept-level error handling in spoken dialogue systems. In *Proc. SIGdial Workshop on Discourse and Dialogue*, pages 178–189.

Janienke Sturm, Els den Os, and Lou Boves. 1999. Issues in spoken dialogue systems: Experiences with the Dutch ARISE system. In *Proc. ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*, pages 1–4, Kloster Irsee, Germany.

Ming Sun, Yun-Nung Chen, and Alexander I. Rudnicky. 2015. Learning OOV through semantic relatedness in spoken dialog systems. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1453–1457.

Stefan Ultes. 2019. Improving interaction quality estimation with BiLSTMs and the impact on dialogue policy learning. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 11–20.

Stefan Ultes and Wolfgang Minker. 2014. Interaction quality estimation in spoken dialogue systems using hybrid-hmms. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 208–217.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proc. Annual Meeting of the Association for Computational Linguistics and Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL)*, pages 271–280.

Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge base completion via search-based question answering. In *Proc. International Conference on World Wide Web (WWW)*, pages 515–526.

Jason Weston. 2016. Dialog-based language learning. In *Proc. International Conference on Neural Information Processing Systems (NIPS)*, pages 829–837.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proc. Conference on Artificial Intelligence (AAAI)*, page 3351–3357.

Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. 2018. Augmenting end-to-end dialog systems with commonsense knowledge. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 4970–4977.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4623–4629.