Dialogue & Discourse 13(1) (2022) 96–122 doi: 10.5210/dad.2022.104

User Impressions of System Questions to Acquire Lexical Knowledge during Dialogues*

Kazunori Komatani KOMATANI@SANKEN.OSAKA-U.AC.JP
SANKEN (The Institute of Scientific and Industrial Research)
Osaka University

Kohei Ono
SANKEN (The Institute of Scientific and Industrial Research)
Osaka University

Ryu Takeda RTAKEDA@SANKEN.OSAKA-U.AC.JP
SANKEN (The Institute of Scientific and Industrial Research)
Osaka University

Eric Nichols E.NICHOLS@JP.HONDA-RI.COM
Honda Research Institute Japan Co., Ltd.

Mikio Nakano MIKIO.NAKANO@C4A.JP
Honda Research Institute Japan Co., Ltd.†

Editor: Barbara Di Eugenio

Submitted 12/2020; Accepted 05/2022; Published online 06/2022

Abstract

We have been working on the challenge of systems that acquire the attributes of unfamiliar terms through dialogues, and we previously proposed an approach based on an implicit confirmation process. The questions posed by a dialogue system must not reduce the user's willingness to converse. In this paper, we conducted a user study that explores the user impression for several question types, including both implicit and explicit questions, to acquire lexical knowledge. User impression scores were collected from 104 participants recruited through crowdsourcing, and a regression analysis was conducted on them. The results demonstrated that implicit questions give a good user impression when their contents are correct, but a bad impression otherwise. The order among the question types combined with their content correctness was also clarified. Furthermore, we found that repeating the same types of questions, even those with correct content, annoys users and lowers the user impression. Our results provide helpful insights for avoiding degradation of user impression during knowledge acquisition.

Keywords: dialogue system, knowledge acquisition during dialogue, lexical acquisition, user impression

1. Introduction

Recently, considerable attention has been paid to *non-task-oriented* dialogue systems in research and commercial system development (Higashinaka et al., 2014; Yu et al., 2016; Smith et al., 2020; Nakano and Komatani, 2020; Roller et al., 2021). In addition to pure non-task-oriented systems,

*. This paper is a modified and extended version of our earlier report (Komatani and Nakano, 2020).
†. Currently, C4A Research Institute, Inc.

certain task-oriented dialogue systems can also engage in non-task-oriented dialogues (Lee et al., 2009; Dingli and Scerri, 2013; Kobori et al., 2016; Papaioannou and Lemon, 2017) because such dialogues are expected to build *rapport* (Bickmore and Picard, 2005; Lucas et al., 2018) between users and systems. Because building an open-domain, non-task-oriented dialogue system that always produces appropriate utterances is a difficult task (Smith et al., 2020; Roller et al., 2021), we consider it worthwhile to build a closed-domain, non-task-oriented dialogue system that attempts to continue dialogues in a specific domain for the purpose of interaction itself.

Ideally, a closed-domain, non-task-oriented dialogue system should have comprehensive knowledge within its domain, such as lexical knowledge. In reality, all such knowledge cannot be prepared in advance. A knowledge base is not only necessary for providing various services such as information search and recommendation, but also effective for non-task-oriented dialogue systems in order to prevent 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). Accordingly, we must consider the case in which a human user uses terms[1] outside the system's vocabulary, i.e., new terms whose ontological categories are unknown to the system.

One of the most important features of a dialogue system is the ability to acquire knowledge from users and so expand its knowledge base through dialogues. Although knowledge may be obtained by asking users to input information into graphical user interfaces (GUIs) or spreadsheets, knowledge acquisition through dialogues beneficially allows the users to enjoy conversations with the system, especially when the system can engage in non-task-oriented dialogues (Kobori et al., 2016). One target of knowledge acquisition through dialogues is obtaining an attribute of an unknown term by asking an appropriate question. This process, here called *lexical acquisition*, allows systems to keep learning even from dialogues containing unknown terms (Meena et al., 2012; Sun et al., 2015). Life-long learning (Chen and Liu, 2018) is an emerging topic that started with machine learning tasks and involves ongoing improvement in a classification, for example. Lexical acquisition during dialogues is a type of life-long learning that can potentially become a key technology for the autonomous evolution of intelligent systems.

A dialogue system can easily ask questions, but repetitive questions can damage the user experience. A dialogue has to continue to allow a system to acquire a variety of knowledge, but users might stop interacting with a dialogue system that repeatedly asks annoying questions. The users are not crowdworkers who repeatedly tell a system whether a target is correct or wrong (Amershi et al., 2014). Instead, questions for knowledge acquisition should be designed to avoid excessive irritation to users. Question design that does not annoy users is an important component when developing non-task-oriented dialogue systems. Consolidating this fact, the Alexa Prize Socialbot Grand Challenge has made conversation duration a vital criterion along with user ratings (Fang et al., 2018; Yu et al., 2019; Finch et al., 2020).

To acquire domain knowledge with less annoying questions, our previous approach adopts an *implicit confirmation* process (Ono et al., 2016), in which a dialogue system produces an utterance about an estimated attribute and determines the correctness of that attribute from the user's response and other contextual information. Figure 1 shows an example of this process. First, when an unknown term emerges in a user's utterance, the system estimates an attribute of that term (Otsuka et al., 2013) (Step 1). Second, instead of asking an explicit question, the system makes an implicit question about the estimated result (Step 2). The implicit question is not a direct interrogative

1. By *term*, we mean an expression denoting an entity that can exist in a knowledge base and may comprise multiple words.

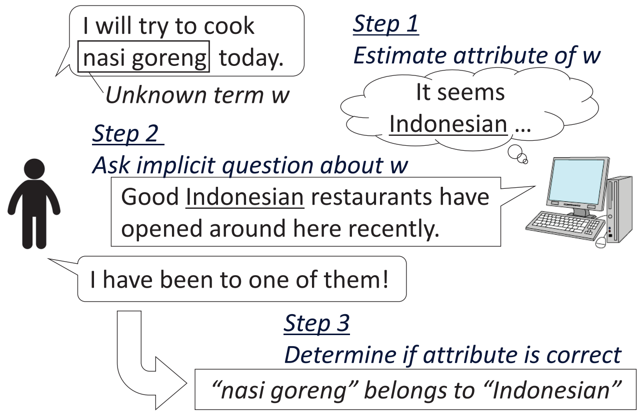

Figure 1: Example of an implicit confirmation process.

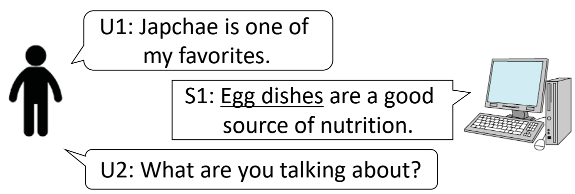

Figure 2: Example of an implicit question with wrong content.

statement; rather, it operates as a question by interpreting the subsequent user utterance together. Third, the system determines the correctness of the estimated result in the implicit question by accounting for the subsequent user response (Step 3). If the estimated result is correct, it is added to the knowledge base of the system.

To enable dialogue systems to acquire knowledge through dialogues while reducing user discomfort, both implicit and explicit questions should be issued in accordance with the situation. For this purpose, we need to investigate when the implicit confirmation process provides a better user impression than asking an explicit question. Moreover, because the questions examine estimation results (in this study, the cuisine types of food names), they can contain wrong content.[2]

The main contribution of this paper is the results of a user study that explores the impact of knowledge acquisition questions on users' impressions. We addressed two specific research questions (RQs):

RQ1 How do the system's question types affect user impressions?

RQ2 Are consecutive explicit questions for knowledge acquisition more annoying to users than consecutive implicit questions?

2. As will be briefly presented in Section 3, we previously tackled the problem of determining the correctness of an estimated result in a question asked during the implicit confirmation process (Ono et al., 2017). For example, Figure 1 shows an example of an implicit question with correct content. The confirmation process appears to be smooth and the question appears not to bother the user. In contrast, Figure 2 exemplifies an implicit question with wrong content, i.e., japchae is not an egg dish but a particular Korean dish. Such questions with wrong content may annoy the user.

User impression data were collected by an experiment in which crowdsourced workers participated in dialogue sessions, which involved three short interactions with a dialogue system. The user impression data were then regressed against the question types used in the session. To answer RQ1, we employed five types of questions including explicit and implicit questions, as well as correct and wrong content. Because the estimated results used in the questions are not always correct, we need to consider the impact of such wrong content on users' impressions. To answer RQ2, we compared the average user impression scores when the same question types were actually repeated during the data collection with the predicted scores of such cases in the regression model.

The remainder of this this paper is organized as follows. After reviewing related work in Section 2, we describe the implicit confirmation process based on our previous experiments (Ono et al., 2017) in Section 3 and determine the correctness of the content in implicit questions. The main contribution of this paper is given in Section 4, which describes the user's impression of various implicit and explicit question types. Section 5 concludes the paper and discusses future work.

2. Related Work

2.1 Knowledge Acquisition in Dialogue Systems

Computers that continually acquire their own knowledge have been long desired. 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 techniques developed for machine learning tasks (such as information extraction) can continuously enhance the performance of classifiers in a semi-supervised manner. This learning process is known as life-long learning (Chen and Liu, 2018). We aim to create systems that acquire knowledge through dialogues with users.

Knowledge acquisition by dialogue systems has been reported in a number of studies. In Meng et al. (2004) and Takahashi et al. (2002), lexical information in dialogues was gained by methods that place unknown terms into coarse categories that generally equate to named entity categories. The coarse categories can be acquired more easily than the more specific categories sought by our approach. The method of Holzapfel et al. (2008) enables a robot to acquire fine-grained categories for unknown terms by iteratively asking questions. However, as this strategy repeats explicit questions, it is unlikely to be appropriate for non-task-oriented dialogue systems.

Pappu and Rudnicky (2014) designed strategies for asking questions in a goal-oriented dialogue system and analyzed the acquired knowledge through a user study. Hixon et al. (2015) proposed a method that poses questions to users and obtain the relations between concepts in a question-answering system. Weston (2016) designed 10 tasks and demonstrated that supervision through feedback from simulated interlocutors improves the utterance-prediction ability of an end-to-end memory network. Li et al. (2017) indicated that asking questions improves the performance of a system employing Weston's method with reinforcement learning. Mazumder et al. (2019) proposed a dialogue system that asks questions about a triple by using the knowledge graph completion.

In these contexts, favorable user impressions of the system's questions are essential for maintaining the dialogues and allowing the system to acquire a variety of knowledge.

2.2 Relationship with Implicit Confirmation Requests in Task-Oriented Dialogues

An implicit confirmation request is a well-known error handling technique for task-oriented spoken dialogue systems (Bohus and Rudnicky, 2005; Skantze, 2005). Many researchers conducted studies to change the form of the confirmation requests, including explicit and implicit ones (Bouwman et al., 1999; Komatani and Kawahara, 2000). For example, consider a flight reservation system that attempts to determine the destination of a user wishing to travel to Seattle. The system can explicitly ask the user "Are you going to Seattle?" then continue the dialogue by implicitly asking "To get to Seattle, where will you depart from?" Previous research has shown that an implicit confirmation request can reduce the number of turns when the content is correct, but correcting the system's misunderstanding when the content is incorrect is difficult (Sturm et al., 1999). In other words, implicit confirmation requests involve a tradeoff between conversation fluency and the risk of taking longer time to make corrections. Most of the contemporary spoken dialogue agents accommodate this tradeoff; explicit confirmations depend on speech recognition confidence scores (Pearl, 2016).

Implicit questions for non-task-oriented dialogues have different goals from those of implicit confirmation requests for task-oriented dialogues. Unlike task-oriented dialogues, implicit questions in non-task-oriented dialogues do not attempt to reduce the number of turns; rather, they lessen the risk of irritating the user and consequently quitting the dialogue. To enable ongoing dialogues between non-task-oriented systems and actual users, we must investigate user impressions of these systems, notably, the acceptability of a certain question type. In particular, the user's impression of non-task-oriented dialogues must not be impaired. However, questions aimed at knowledge acquisition in non-task-oriented dialogues have been rarely discussed.

2.3 User Satisfaction and Impression in Dialogues

Several studies have focused on predicting user satisfaction of dialogues. Walker et al. (1997) proposed a methodology that predicts user satisfaction in task-oriented dialogues using a regression model with several objectively obtainable parameters. Using a hidden Markov model, Higashinaka et al. (2010) modeled user-satisfaction transitions even when given only the ratings of entire dialogues. Ultes and Minker (2014) and Ultes (2019) employed different context-aware machine learning methods to improve the prediction accuracy of interaction quality, which is defined similarly to user satisfaction.

Our goal in this paper is *not* to predict user satisfaction for dialogue evaluation but rather to analyze the effects of various question types on users' impression in dialogues. In non-task-oriented dialogues, user impressions can be regarded as comparable to user satisfaction because the user is expected to enjoy the dialogue. This situation differs from task-oriented dialogues, in which user satisfaction depends on the task success and dialogue cost Walker et al. (1997). Kageyama et al. (2018) investigated whether gradually controlling the form of a system's utterances can improve users' impression.

Rather than assessing a whole dialogue system, we focus on one component of user satisfaction, namely, the user's impression of the system's questions. Specifically, we explore the users' impression of diverse question types, including explicit and implicit questions, for knowledge acquisition during non-task-oriented dialogues. Maintaining high user impressions is vital in a non-task-oriented dialogue system that strives to acquire knowledge from users, because users will cease interacting with a system that repeatedly asks irritating questions.

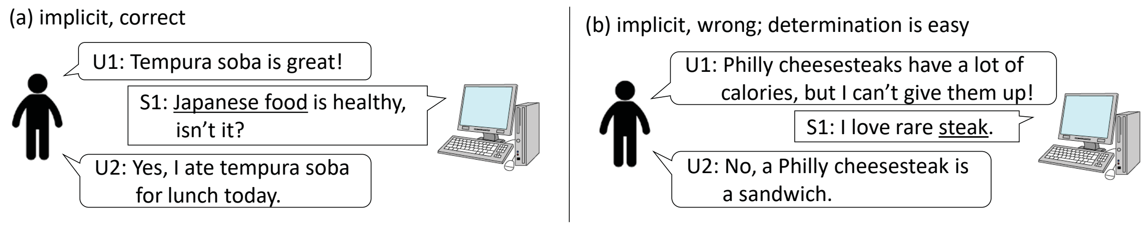

Figure 3: Examples of correct and wrong implicit questions.

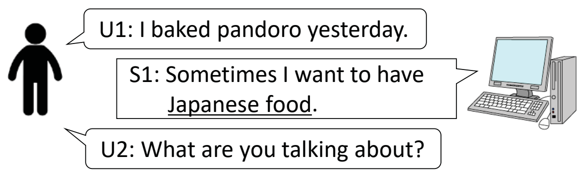

Figure 4: Example of an implicit question for which a wrong attribute is not easily determinable.

3. Implicit Confirmation Process of Knowledge Acquisition

This section discusses the concept of the implicit confirmation process. This process presents a possible approach to acquire knowledge through dialogues in non-task-oriented dialogues, but it has received little attention.

Previously, we showed that the implicit confirmation process can be used to determine the correctness of an estimated attribute of an unknown term (Ono et al., 2017) (Step 3 in Figure 1). For example, Question `S1` in Figure 3 (a) does not explicitly ask the user whether “Japanese” is an attribute of the dish *tempura soba*, but the user response `U2` informs the system that the attribute is correct. Figure 3 (b) shows another example in which the system determines from `U2` that the estimated attribute is wrong. Because this approach was promising in our earlier study, the implicit confirmation process can plausibly be applied to knowledge acquisition in our present study. Users' impressions of this process, which have thus far remained unanswered, will be presented in Section 4. This section introduces the problem and discusses the viability of the approach.

In an implicit confirmation process, whether an estimated attribute is correct cannot always be determined. Because implicit questions can elicit many different forms of responses, simply examining the linguistic expressions of a user's responses is insufficient. For example, in Figure 4, the system wrongly estimates the attribute *Japanese food* for *pandoro* mentioned in `U1`, although pandoro is Italian. The system then forms an implicit question `S1`. In such cases, the estimated attribute is not easily detected as wrong because the user utterance `U2` includes no literally negative expressions, such as “no” and “not.” To avoid these problems, we require a method that precisely determines the correctness of the estimated attributes through the implicit confirmation process.

In some studies, affirmative and negative sentences are classified using rules or statistical methods. For example, de Marneffe et al. (2009) defined rules for judging whether a response to a yes/no question is affirmative or negative when the response is not a simple “yes” or “no.” Gokcen and de Marneffe (2015) investigated features for detecting disagreement in a corpus of arguments on the

Table 1: Features of binary classification with user responses.

f1	U2 includes an affirmative expression in response to S1
f2	U2 includes a literally negative expression in response to S1
f3	U2 includes an expression correcting S1
f4	U1 and U2 contain the same term
f5	U2 includes the category name in S1
f6	U2 includes another category name other than that in S1, excluding cases that fall under f3
f7	U2 includes a word preventing a change in topic in S1
f8	U1 includes the category name in S1
f9	U1 includes another category name other than that in S1
f10	U1 includes any interrogative
f11	U1 includes an expression corresponding to the category mentioned in S1

Web. Such outcomes are helpful for interpreting user responses to explicit questions. In contrast, we attempt to determine the correctness of an attribute in an implicit question not only from the user response but also from the surrounding context in the implicit confirmation process.

3.1 Determining the Correctness of the Content of an Implicit Question

Based on the implicit question and the previous and succeeding user utterances, we determine the correctness of the estimated category in an implicit question. Determining whether an estimated attribute is correct or wrong in an implicit confirmation process can be cast as a binary classification problem. Accordingly, an experiment was conducted by using collected data with user responses. We tested the classification performance between the binary labels "correct" and "wrong." The classification result can then be used to determine whether the system should add the term-category pair to its knowledge base.

We presume that the system can detect a food name in the user's input, by using methods such as named entity recognition (Mesnil et al., 2015), even when that name is outside the system's vocabulary. The ontological category of an unknown term is an attribute to be estimated. We further assume that a category can be estimated with an existing method, as in Otsuka et al. (2013). We make no assumptions about the ontological structure for food.

3.2 Evaluation of the Correctness of the Estimated Category

3.2.1 FEATURES FOR THE CLASSIFICATION

Table 1 lists the features for classifying the correctness of categories in implicit questions. In the above figures, U1, S1, and U2 respectively represent a user input, the system's implicit question after U1, and the user response to that question. All feature values are binary; if the statement of each feature is true in a given circumstance, that feature takes the value 1; otherwise, its value is 0. These features were designed to represent differences in the expressions of user responses to implicit questions with either a correct or wrong category.

The following expressions were manually prepared for each feature (in Japanese, the language in which the experiment was conducted). These expressions were used in the present experiment, but they can be expanded to those that appear in the data. Lexical variations can be handled by current techniques enabling more sophisticated use of word embeddings.

Feature f1 represents the situation when a user responds an affirmative expression in response to an implicit question with a correct category. Affirmative expressions for this feature included "yes," "that's right," and 13 other expressions. Similarly, feature f2 represents the situation of negative expressions, which tend to be used for wrong implicit questions. Literally negative expressions for this feature included "is not the [category name used in `S1`]", "no", and 15 other expressions. In this paper, features f1 and f2 are employed as the baseline condition because they consider only affirmative and negative expressions in `U2`, ignoring the relationship between `U2` with `S1` or `U1`.

When the system asks an implicit question in a wrong category, the user tends to perceive that the system has abruptly altered the topic. Feature f3 attempts to detect this situation to correct the system's previous question `S1`, with one of six expressions in `U2`; for example, "it is [another category name other than that in `S1`]."

Feature f4 represents the situation in which a category in an implicit question is correct and the user carries the topic in `U1` into `U2`, as shown in Figure 3 (a). Feature f4 aims to capture this situation by detecting the same term in `U1` and `U2`: *tempura soba* in this example.

Similarly to Feature f3, Features f6 and f7 attempt to detect when the user's topic in `U2` differs from that in `U1`. For example, consider the following example:

> `U1`: I like sangria with its fruity taste.
> `S1`: *Yogashi* has a rich taste, doesn't it?
> `U2`: I am talking about the alcoholic beverage.

In `S1` of this example, the system asks an implicit question with the wrong category *yogashi*[3]; the correct category of sangria is *alcoholic beverage*. Subsequently, the user tries to return to the original topic of alcoholic beverage. `U2` includes a category name other than that in `S1`. This situation is represented by feature f6, which excludes cases falling under feature f3 to maintain exclusivity of the two features. In such cases, `U2` often contains the Japanese word *hanashi*[4], which is represented as feature f7. The present experiment considered only one word.

Features f6 and f9 covered 25 category names: 20 categories included in the system's implicit questions, such as "Japanese food," "Italian food," and "Korean food", and five food names such as "cheese" and "pasta." Feature f10 covered 18 interrogative expressions. Feature f11 checks for discrepancies between the expressions "eat" and "drink" in `U1` and the categories contained in `S1`. Specifically, when U1 contains "eat" or a conjugated form of "eat", the feature value depends on whether or not `S1` contained a category related to "drink" (e.g., an alcoholic beverage), and vice versa. This feature may be domain-dependent because it assumes that each category corresponds to either "eat" or "drink."

3.2.2 DATA AND SETTING

Twenty pairs of terms and corresponding implicit questions were prepared for the experiment. Ten of the categories in the implicit questions were correct; the remaining ten were wrong. For example, an implicit question relating to *churrasco* with its correct category *meat dish*[5] was "Eating meat is fun, isn't it?" As another example, an implicit question related to *sangria* with a wrong category

3. Yogashi means western sweets in Japanese.
4. For instance, this word tends to be used to say "I am talking about ..." and "What are you talking about?" in Japanese. Feature f7 may be unique to the Japanese language.
5. Note that food category hierarchies in Japan may differ from those in other countries.

Table 2: Confusion matrices.

		Reference	
Features	Output	Correct	Wrong
all	Correct	742	313
	Wrong	236	665
f1, f2 only	Correct	320	220
	Wrong	658	758

Table 3: Classification results.

Features		Precision	Recall	F-score
All	Correct	0.703	0.759	0.730
	Wrong	0.738	0.680	0.708
f1, f2 only	Correct	0.593	0.327	0.422
	Wrong	0.535	0.775	0.633

yogashi was "Yogashi has a rich taste, doesn't it?" Furthermore, the phrases of the implicit questions were subtly tweaked to improve their naturalness when the user's input was interrogative or negative.

Data were collected on 1,956 responses from 98 workers through crowdsourcing. The data were evenly split between the responses to implicit questions with correct and wrong categories. The data from two workers who input only the specified terms or repeated the same sentences were removed. Four invalid inputs containing spaces only were also eliminated.

Classification was performed by logistic regression[6]. Specifically, we used the Weka module (version 3.8.1) (Hall et al., 2009) with its default parameter settings. The classification results were evaluated using a 10-fold cross-validation.

3.2.3 RESULTS

The results of two feature sets were compared in the experiment: one including all 11 features listed in Table 1 and the other including a baseline set consisting only of features f1 and f2.

Table 2 presents the results (confusion matrices) of raw outputs of both feature sets

The classification accuracies on the complete and baseline feature sets were 71.9% (1,407/1,956) and 55.1% (1,078/1,956), respectively. As confirmed in a McNemar test, this difference was statistically significant ($p < .001$), affirming that incorporating the features expressing context improved the classification performance over using the features obtained from `U2` alone.

Feature selection also revealed the most discriminant features for the classification. Specifically, the experiments were repeated for the 11 features, i.e., $2,047$ $(= 2^{11} - 1)$ feature sets, and the average F-scores of the combinations were compared. Table 3 summarizes the precision, recall, and F-scores of the two classes ("correct" and "wrong"). When using all features and f1 and f2 alone, the average F-scores, i.e., the arithmetic means of the F-scores of the two classes, were 0.719 and 0.528, respectively.

Table 4 lists the top 10 feature sets ranked by their average F-scores. The condition "None" under which all 11 features were used ranked second in the table, indicating that almost all features effectively contributed to the classification. After eliminating f10, the F-score of the "wrong" cate-

6. A more modern classifier, such as a Transformer-based classifier, might improve the classification performance.

Table 4: Top-10 feature sets after removing the arbitrary features for classification.

Removed features	Correct P	Correct R	Correct F	Wrong P	Wrong R	Wrong F	Avg-F
f10	.704	.759	.730	.738	.681	.709	.719
None	.703	.759	.730	.738	.680	.708	.719
f7,f10	.701	.760	.729	.738	.676	.705	.717
f1,f4,f10	.699	.764	.730	.740	.672	.704	.717
f1,f4	.699	.765	.730	.740	.671	.704	.717
f7	.701	.759	.729	.737	.676	.705	.717
f4,f10	.691	.784	.735	.751	.649	.696	.715
f4	.690	.784	.734	.750	.648	.696	.715
f1,f4,f7,f10	.696	.765	.729	.739	.666	.700	.715
f1,f4,f7	.695	.766	.729	.739	.665	.700	.715

P: precision, R: recall, F: F-score

gory marginally improved, leading to an improvement in the overall F-score. Because f10 appears numerous times in the table, it was less helpful for the classification than other features. When f10 was removed, f8 had the highest positive weight in the logistic regression function, indicating that f8 gave strong evidence for the "correct" category when its value was `1`. Consequently, when a category name appeared in both `U1` and `S1`, the category in `S1` was likely to be correct because the topic was not suddenly shifted.

4. User Study to Investigate Users' Impression of Questions

We then explored users' impressions of implicit and explicit questions. Specifically, two specific research questions were addressed. First, we determine the impact of the system's question types on user impressions. Second, we evaluate whether consecutive explicit questions for knowledge acquisition are more annoying than consecutive implicit ones.

The data collection was designed to satisfy the following conditions: (1) the user should not be excessively annoyed by the process (2) Any effect of the consecutive explicit questions should be discernible We thus adopted a design in which a question survey followed after several sub-dialogues (three in this paper) were repeated as a session, (see Figure 5). Although the user's impressions could simply be assessed after every system question, this design would be extremely inconvenient and break the dialogue flow. Instead, we quantified the influence of each question type in one session using a regression model, which also provided an analysis of user impressions after repeating the same question type.

4.1 User Study Setting

We assumed a dialogue system that asks an attribute value for an unknown term. In other words, when an unfamiliar term arises in a dialogue, the system attempts to acquire its attribute from the user through the dialogue. The term and its attribute pair can then be stored as new system knowledge.

	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 5: Five question types of *puttanesca*, whose correct cuisine type is *Italian*, with examples. E and I respectively denote explicit and implicit questions, C and W respectively denote whether the estimated cuisine is correct or wrong, and Whq denotes a wh-question.

In the present experiment, we assumed that an unknown food name can be paired with its cuisine type. First, the cuisine type was estimated from the food name's character sequence (Otsuka et al., 2013). The estimated cuisine was then verified by asking a question.

4.1.1 FIVE QUESTION TYPES FOR KNOWLEDGE ACQUISITION

Table 5 lists examples of the five question types. In these examples, the unknown term is *puttanesca*, the estimated correct cuisine is *Italian*, and the estimated wrong cuisine is *Japanese*.

Each question type has two components: the form of the question and the correctness of its content. The first component can be explicit (E), implicit (I), or a wh-question (Whq). An explicit yes/no question asks whether the content of a question is correct (e.g., "Is puttanesca Italian?"). An implicit question continues the dialogue with a system utterance containing the estimated cuisine (e.g., "Italian is perfect for a date"). The system then implicitly determines whether the cuisine is correct by analyzing the subsequent user utterance (Ono et al., 2017). Meanwhile, a wh-question simply asks without any estimated cuisine (e.g., "What is puttanesca?").

The second component is whether the estimated cuisine is correct (C) or wrong (W). This component allows the investigation of whether users' impressions are affected by correct or wrong content resulting from the automatic cuisine estimation of the unknown food name (Otsuka et al., 2013) before the system posed a question. Because wh-questions have no particular content, this component applies exclusively to E and I questions. For simplicity, all questions were one-choice rather than multi-choice (Komatani et al., 2016)

4.1.2 DATA COLLECTION

The data for evaluating user impressions of dialogues including the above-described five question types were collected by another crowdsourcing[7]. All crowdworkers were Japanese speakers and all dialogues were conducted in Japanese. The workers were informed that they would be conversing with an "AI chatbot," and were asked to simulate a first-time conversation with the chatbot.

The workers gave an impression score in each session. The experimental flow is depicted in Figure 5. One session consisted of three sets of interactions, followed by an impression survey.

7. We used the platform CrowdWorks, Inc. (`https://crowdworks.co.jp/`).

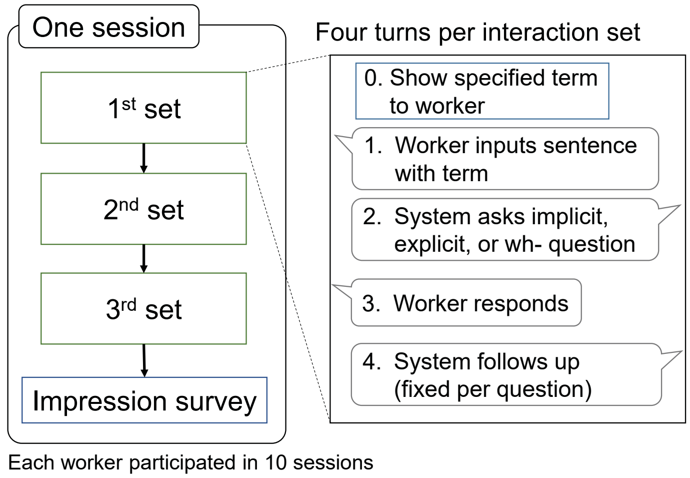

Figure 5: Flow of data collection.

Each interaction set included two system turns and two user turns. Before the first turn, an instruction with a term was displayed, e.g., "Please input your thoughts as though you ate *puttanesca* recently." The flow of the four turns is described below:

Turn 1: The worker types in a sentence including the term specified in the instruction. The terms were prepared before the experiment.

Turn 2: The system asks a question about the term, where the type of the question is randomly selected. Wrong cuisine estimation results and phrases of implicit questions were manually prepared prior to experiment.

Turn 3: The worker provides an unrestricted response to the system question.

Turn 4: The system displays its follow-up response, which depends on the question type[8] selected in Turn 2. That is, the follow-up response was unaffected by the worker's response in Turn 3. For example, when the system's question in Turn 2 was "implicit, wrong (IW)," the system's follow-up response in Turn 4 was always "Sorry, I probably misunderstood."

After three interaction sets, the workers recorded their impressions of the session in the questionnaire shown in Figure 6. The questionnaire used a 7-point Likert scale with two items: "Were the system utterances annoying?" and "Was the system intelligent?"[9] Hereafter, these impression scores are denoted as *annoying* and *intelligent*, respectively.

Each worker was required to participate in 10 sessions. Before the experiment, we manually prepared 30 terms that were unfamiliar to Japanese crowdworkers. These phrases were introduced in fixed order as the unknown terms in each session (three terms per session).

8. Again, the question type is randomly selected and not dependent on automatic estimation, such as determining whether the content is correct or wrong.
9. These questionnaire items were unvalidated, meaning that they did not introduce redundancy (i.e., different ways of asking the same content) to minimize misinterpretations, as suggested in (Davis, 1989). This straightforward approach was chosen because it was simple to communicate to the crowdworkers.

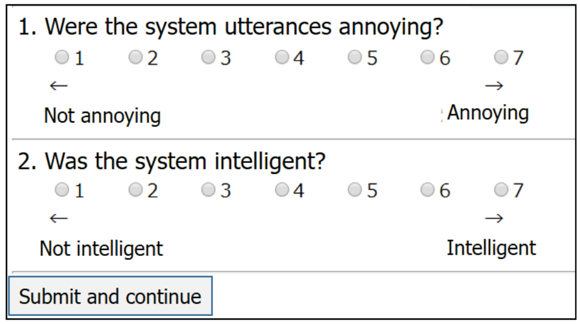

Figure 6: Screenshot of the impression survey.

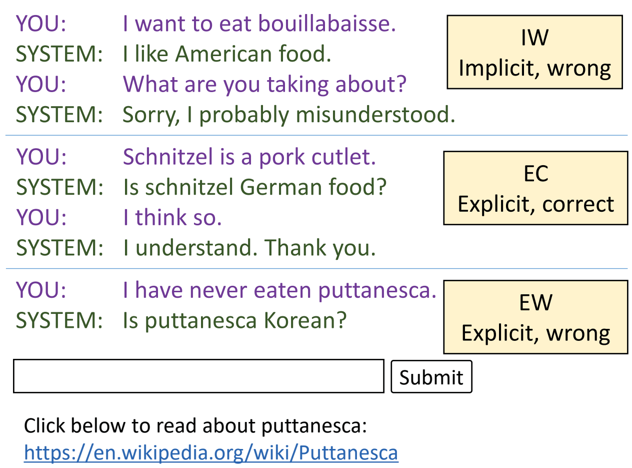

Figure 7: Example of the system image in a session. The boxes on the right enclosing the question types are shown for explanation only and were not displayed to the workers.

Figure 7 is an example of a system image (translated from Japanese). The lines starting with "YOU" and "SYSTEM" denote the utterances of the worker and the system, respectively. The first display of the interaction set, in which the specified term was presented to the worker, is not depicted in the figure because it disappeared after the worker entered the first sentence. If a worker did not recognize a term, a link to Wikipedia was provided at the bottom of the screen for look-up purposes, thereby avoiding dialogues in which the worker was unfamiliar with the meaning of the term. Although the dialogues were not natural, they provide an initial step when (as in the present case) a system that can naturally acquire knowledge through several turns is lacking.

Table 6: Numbers of occurrences of question types in the collected data.

EC	EW	Whq	IC	IW
719	618	650	612	680

Table 7: Summary of the two impression scores obtained on a 7-point scale.

	intelligent	*annoying*
Average	3.812	3.048
Standard deviation	1.562	1.613

Originally, 120 workers collectively completed 1,319 sessions.[10] After removing unusable data (such as data from workers who did not complete all 10 sessions), we obtained a total of 1,093 sessions from 104 workers. That is, we obtained 1,093 *intelligent* and *annoying* impression scores for each session, where each session included three system question types to be analyzed. The numbers of occurrences of the five question types in the collected data are listed in Table 6. These numbers were supposed to be approximately equal but became uneven through the random selection and a system error. The average number of occurrences was 655.8 ($= 1,093$ sessions $\times$ 3 sets $\div$ 5 question types), and the standard deviation was 44.6.

The question type was randomly selected three times from the five types, giving 125 ($= 5^3$) possible question-type patterns for a session. Here, the patterns are represented by concatenating the three question types with hyphens: for example, the pattern in Figure 7 is 'IW-EC-EW'. The actual number of patterns was 124, as one pattern (Whq-Whq-IW) was never chosen by the random selection process. The average occurrence number of each pattern was 8.81 ($= 1,093$ sessions $\div$ 124 question-type patterns), and the standard deviation was 3.96 (maximum: 17; minimum: 0).

Table 7 lists the averages and standard deviations of the two impression scores. The standard deviations were large for a 7-point scale. Because the impression scores were subjective, there was little agreement on scores among the workers, but impression scores in various question types followed a consistent pattern for each worker. It is also worth noting that the trends of the two impression scores were almost opposite, with a Pearson correlation coefficient of -0.512.

4.2 Analysis with Linear Regression

From the regression coefficients of the linear regression model, we extracted the influence of each question type. In the basic model, the explanatory variable was the number of occurrences of the five question types in a session and the objective variable was one of the two impression scores (*annoying* or *intelligent*). The basic regression model for predicting the score of the i-th session was

$$\text{score}_i = w_0 + \sum_{t \in \{EC, EW, Whq, IC, IW\}} w_c \cdot num_i(t), \tag{1}$$

where $num_i(t)$ denotes the number of each question type t used in the session (0, 1, 2, or 3 in the basic model).

To improve the multiple correlation coefficients, we added two improvements to the basic model. First, we normalized the impression scores to obtain a mean of 0 and a variance of 1 for each worker. Normalization eliminates the variations among the workers. The workers recorded a range of impression scores over the 7-point scale; that is, some gave higher scores, while others gave lower scores. To understand the effect of each question type, we used the relative scores given by each worker.

10. Because of a system error, some workers participated in more than 10 sessions.

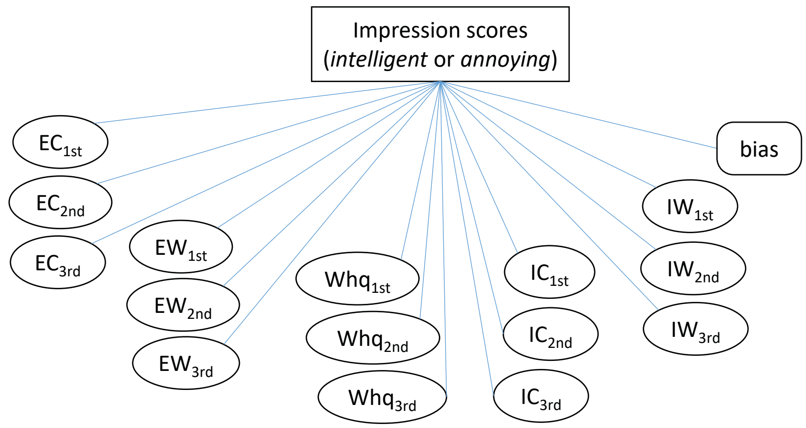

Figure 8: Illustration of the refined regression model. In each of the 15 ovals is a binary value indicating whether a question type occurred at a particular position in the i-th session.

Second, we considered the positions of the questions in each session. This analysis involved 15 independent variables: the five question types times the three positions (representing the first, second, and third interaction sets in a session). Therefore, the refined regression model was

$$\text{score}_i = w_0 + \sum_d w_d \cdot x_{id}, \tag{2}$$

where $d \in \{EC, EW, IC, IW, Whq\} \times \{1st, 2nd, 3rd\}$. The occurrence of each question type d in the i-th session, denoted by x_{id}, takes a binary value (0 or 1) and $\sum_d x_{id} = 3$ for each i. The model represented by Equation (2) is illustrated in Figure 8. The occurrence distributions of the question types among the 15 possible positions in the collected data are ideally equal but were unequal in practice. The average number of occurrences at each position was 218.6 ($= 1,093$ sessions $\times$ 3 sets $\div$ 15 question types and positions), and the standard deviation was 17.4 (maximum: 245; minimum: 196).

The regression coefficients w_d were computed from the collected data using the least-squares approach. These coefficients, which were used in subsequent analysis, represent the change in value of the objective variable (i.e., a user impression score) when each explanatory variable x_{id} is 1, according to causal inference (Angrist and Pischke, 2008). As a precondition of the analysis, each explanatory variable was binary and uncorrelated with any other explanatory variables. Such multicollinearity was avoided because the question types corresponding to the explanatory variables were randomly chosen during the data collection, as detailed in Section 4.1.2.

The regression model for investigating the influence of the explanatory variables (not for predictive purpose) in terms of the coefficients w_d. The resultant w_d values may depend on the collected data and the settings of the explanatory variables from which they were derived. However, because each question type was chosen at random and the values of the explanatory variables were supposedly independent, we believe that the relationship among the explanatory variables has certain generality.

The multiple correlation coefficients for the two impression scores are listed in Table 8. The coefficients increased after normalization and were further increased after accounting for the posi-

	intelligent	*annoying*
Basic regression model	0.368	0.207
+Normalization by worker	0.493	0.308
+Consideration of position	0.540	0.354

Table 8: Multiple correlation coefficients (R) of the models.

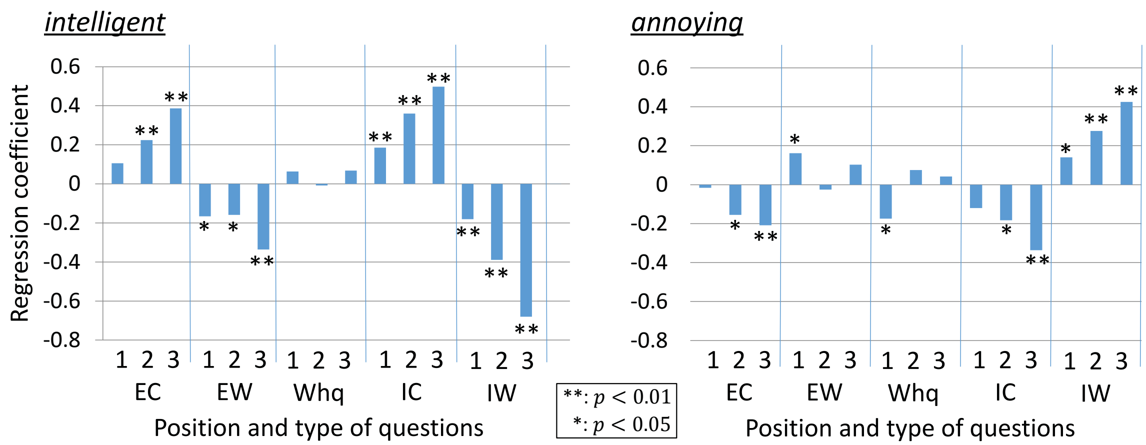

Figure 9: Regression coefficients of the model considering question types and positions.

tions. Accordingly, in subsequent analysis, we employed the refined model after the normalization and position consideration.

4.3 Results

4.3.1 ANALYSIS OF THE OBTAINED REGRESSION COEFFICIENTS

We first consider RQ1: "How do the system's question types affect user impressions?" Figure 9 shows the values of the 15 regression coefficients obtained for the labels *intelligent* and *annoying*. We also checked the statistical significance of the individual regression coefficients being non-zero. The symbols ** and * indicate statistical significance at the $p < 0.01$ and $p < 0.05$ levels, respectively.

In the case of *intelligent*, larger positive values indicate that when the system asked that question type in that position, the workers tended to believe that the system was intelligent. A high positive value thus implies a good impression. In the case of *annoying*, larger positive values imply that when the system asked that question type in that position, the workers tended to believe that the system was annoying. A high positive value thus implies a bad impression.

The averages over the three positions for the two labels are summarized in Table 9. The regression coefficients of the five question types were 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*. Note the opposite orderings of the two impression scores.

	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 9: Average regression coefficients over the three positions.

In the *intelligent* model, the coefficients of IC and EC (implicit and explicit questions with correct content) were positive, whereas those of EW and IW (implicit and explicit questions with wrong content) were negative. In the *annoying* model, the opposite relations held. These results correspond to our intuition that when the system asked questions with wrong content, the workers would regard the system unintelligent and become irritated by its questions. Because the wh-questions had no concrete content, the Whq coefficients were intermediate between those of C and W. However, the Whq coefficient for *annoying* was small and negative, implying that the first wh-questions were not particularly annoying.

We now explore the relationship between the explicit and implicit questions. The absolute values of the regression coefficients of the IC questions were larger than those of the EC questions. This result suggests that the implicit questions tended to give a better impression than the explicit ones. As a reason of this trend, we suggest that the workers perceived a knowledge of rare and difficult terms by the system. Specifically, the impression scores were higher for target foods with uncommon names that for well-known foods. In contrast, the absolute values of the coefficients of the IW questions were larger than those of the EW questions. In other words, when the estimated cuisine was wrong, the implicit questions gave a worse impression than the explicit ones. In this case, the workers probably perceived that when the system implicitly asked about the wrong cuisine, it had ignored the user's previous utterances and had switched the dialogue to a new topic.

Figure 9 also reveals the tendencies among the three positions for each question type. In the case of *intelligent*, the negative and positive regression coefficients of all five types were largest at the third position. In the case of *annoying*, the negative and positive regression coefficients of question types EC, IC, and IW were largest at the third position. Therefore, the type of question asked soonest before the impression survey significantly affected the impression scores.

4.3.2 IMPRESSION OF REPEATING THE SAME QUESTION TYPE

We next consider RQ2: "Are consecutive explicit questions for knowledge acquisition more annoying to users than consecutive implicit questions?" In this study, we compared the following two impression scores:

- Actual scores when the same question type was asked three times.
- Scores predicted by the regression model.

The former scores were calculated by averaging the scores of the sessions in which the same question types were actually repeated through random selection. Because the question type was randomly chosen during the second turn of a session (see Section 4.1.2), the probability of selecting the same question type three consecutive times was $1/5^3$. In the collected data, such occurrences averaged 10.4 times per question type.

The latter scores were calculated using the model of Eq. (2) in the virtual case of selecting the same question type three consecutive times. e.g., by substituting $num_i(\mathrm{EC}_{1st}) = num_i(\mathrm{EC}_{2nd}) =$

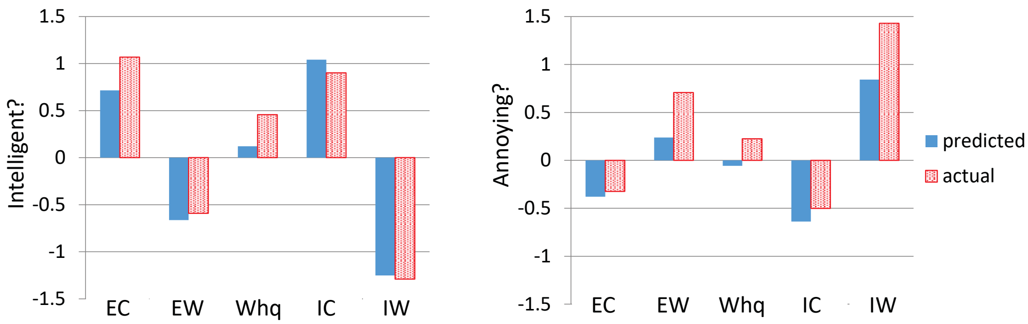

Figure 10: Impression scores (predicted vs. actual) for *intelligent* (left) and *annoying* (right).

	intelligent			*annoying*		
	Predicted	Actual	Difference	Predicted	Actual	Difference
EC	0.716	1.068	+0.352	−0.380	−0.324	+0.056
EW	−0.663	−0.591	+0.072	0.239	0.707	+0.468
Whq	0.122	0.457	+0.335	−0.058	0.224	+0.282
IC	1.042	0.901	−0.141	−0.639	−0.502	+0.137
IW	−1.251	−1.290	−0.039	0.842	1.429	+0.587

Table 10: Predicted and actual impression scores and their differences when the same question type was repeated three times.

$num_i(\text{EC}_{3rd}) = 1$ in Eq. (2). Because the coefficients were calculated from data in which each question type was randomly chosen, the predicted scores represent cases when the five question types appeared in various contexts.

By comparing the two scores, we can analyze the impact of consecutive questions by the average impression scores when the same question type was *actually* repeated and when a question type appeared in various contexts with those of the latter case represented by the *predicted* scores.

Figure 10 shows the comparison and Table 10 lists the actual values. We first investigate the results of *annoying* (right panels in the figure and right columns of the table) because the goal of RQ2 was to check the displeasure level of asking consecutive questions. For all question types, the impression scores were larger in the actual cases than in the predicted cases, indicating that asking repeated question types was more annoying. Furthermore, the scores for the EW and IW questions showed much wider differences than those for the EC and IC questions, as seen in the "Difference" column in the right part of Table 10. The finding is consistent with our intuition that asking a series of questions with wrong content is more annoying than asking a series of questions with correct content.

Examining the results for *annoying*, we observe that in both the predicted and actual cases, the scores of the IC questions were lower than those of the EC questions, indicating that even when the contents are correct, implicit questions are less annoying, than the explicit questions. The order between the EC and IC questions did not alter, although the degrees of "annoying" were larger in the actual cases than in the predicted cases. When the content was correct, consecutive EC questions were more annoying than consecutive IC questions. Therefore, the answer to RQ2 is affirmative

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 *bagna cauda*. It was like shabu-shabu.
System	Is bagna cauda Italian?
User	Yes. It is an Italian pot cuisine.
System	I understand. Thank you.

Figure 11: Example of a session during which explicit questions with correct content (EC questions) were asked three consecutive times.

when the questions have correct content. One possible explanation for this outcome is nearly identical forms of the explicit questions (e.g., "Is puttanesca Italian?"), giving the impression that the system repeated the questions without thought. In contrast, implicit questions caused less disruption and were therefore less annoying because they retained the content of the worker's previous utterance when the cuisine was correctly estimated.

The results for *intelligent* are presented in the left panel of Figure 10 and the left columns of Table 10. The impression scores for *intelligent* in the actual cases varied with question type. The impression scores increased when the system actually asked an EC-type question (i.e., explicit question with the correctly estimated cuisine) three times. We assume that the workers evaluated the system as intelligent because it understood the cuisines of unfamiliar food names. Meanwhile, the impression scores for EW and IW did not alter significantly, even when the system actually posed questions with wrong content three times. We speculate that the workers perceived the system as unintelligent, once the system posed a question with wrong content. When the system actually posed an IC-type question three times, the drop in the results might be explained by the invariant phrasing of the IC questions, giving the impression of a monotonous and unintelligent system. When wh-questions were actually asked three times, the scores for Whq increased, indicating that the workers regarded the system as intelligent despite being irritated by the questions (see results for *annoying*). Note that for *intelligent*, the results of the Whq questions were lower than in EC and IC cases.

Figures 11 and 12 show example dialogues from sessions during which EC and IC questions, respectively, were asked three times.

4.4 Discussion

We clarified the order of the question types as either *intelligent* or *annoying*. Even when the content is correct, we found that repeating the same question type annoys users and lowers their impressions of the system. As long as the content is correct, implicit questions tend to give a better user impressions than explicit ones. In contrast, when the content is wrong, implicit questions are viewed less favorably than explicit ones.

These experimental results justify an adaptive strategy based on the cuisine estimation results to lower the risk of degrading the users' impression. Specifically, a dialogue system should select

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 flavors.
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 12: Example session in which implicit questions with correct content (IC questions) were asked three consecutive times.

an implicit question if the confidence score of the cuisine estimation is high and an explicit one (or a wh-question) if the confidence score is low. This adaptive strategy aligns with the prior findings on task-oriented dialogues using confidence measures of automatic speech recognition (e.g., (Sturm et al., 1999; Pearl, 2016)), as discussed in Section 2.2. Incorporating the utility of each question type for knowledge acquisition (Komatani et al., 2016) would be an interesting extension of this strategy.

The analysis results discussed in Section 4.3.2 confirmed the earlier best practice of the designers of dialogue systems: that is, the system must avoid repeating the same type of questions in non-task-oriented dialogues. Instead, the system should contain multiple question types to engage in smooth dialogues with users. Question types should be appropriately changed by considering not only the confidence of estimated cuisines but also the history of the dialogue. The system can effectively acquire knowledge through such dialogues and continue the dialogues without downgrading the user's impression.

Varying the question phrases is also worth of exploration. The set expressions of our present experiment might have imparted a monotonous, annoying impression to users. We would therefore be interested in the outcome of syntactically altering the phrases of explicit questions. As the phrases of the implicit questions were likewise fixed for the estimated categories, we are similarly interested in enhancing the diversity of implicit expressions.

5. Conclusion

Through a user study, we addressed a key issue in the implicit confirmation process (Ono et al., 2016) of non-task-oriented dialogue systems: whether implicit questions and explicit questions elicit different user impressions of the system. The user impressions were investigated on five types of questions. We clarified the order and found that even when the content is correct, repeating the same question type irritates users and lowers their impression of the system.

Implicit confirmation is a promising question strategy for a non-task-oriented dialogue system that attempts to acquire more knowledge through dialogue without bothering the users with simple, repeated, explicit questions. The presented findings and methodology will be useful for analyzing how different question types influence user impressions and for designing questions for a system that acquires knowledge effectively through dialogues with users.

Several issues should be resolved in future work. The number of turns and domain of our experiments were constrained. To remove these constraints, we must evaluate non-task-oriented dialogue systems that can engage in longer dialogues in several domains. We are planning to implement a non-task-oriented dialogue system that can acquire knowledge via an implicit confirmation process embedded within a longer dialogue. The process can be implemented by preparing expressions of implicit questions for each category to be estimated (cuisine types in the current study). This implemented system will be tested in another user study.

The present study was not performed in a specific context or situation. This crucial factor must be considered in future work. We discussed knowledge acquisition during non-task-oriented dialogues (e.g., chatting about food), but when a user is teaching the system, the system will be allowed to ask questions repeatedly. The user's motivation to talk with the system will also change according to a situation. Nonetheless, the present findings indicate how the system can prevent the user from losing motivation in continuing the dialogue. In particular, the system must pick appropriate questions and thereby avoid degradation of the user's impression.

An essential problem in knowledge acquisition is that users' responses may differ, for example, some users may say that *mapo doufu* is Sichuan, while others may claim it is Chinese. This difference arises from the different granularity degrees of users' ideas, as evidenced in their responses. A knowledge graph with different nodes representing such concepts might resolve this challenge. We could also use confidences on the correctness of the question content given by knowledge graph completion results (Komatani et al., 2021).

Extending the findings to spoken interactions is another interesting avenue as it necessitates the use of automatic speech recognition. Phoneme recognition techniques, which convert speech signals to phonetic symbols, would be helpful to handle unknown terms. The recognition accuracy of these techniques has been improved by acoustic models based on deep neural networks, but at least two key difficulties remain: (1) The segmentation of recognized phonetic symbols into words, and (2) the distinction between unknown and misrecognized terms. The former difficulty has been tackled by approaches based on Bayesian models (Heymann et al., 2014; Takeda et al., 2018; Takeda and Komatani, 2019). The latter problem relates to misspelled words in text inputs. Distance metrics between an input and known terms are potentially useful for identifying unknown "long" terms rather than "short" ones, where long and short refer to the numbers of the phonetic symbols in the term. We would need to identify unknown terms through dialogue, as no known solution can completely prevent segmentation and recognition failures. Acquisition of knowledge, especially of unknown terms, through spoken dialogues still requires favorable user impressions and the dialogue strategy will play a key role in maintaining users' motivation to continue the dialogues with the system.

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. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4), December 2014. URL https://doi.org/10.1609/aimag.v35i4.2513.

Joshua Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

Timothy W. Bickmore and Rosalind W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2): 293–327, June 2005. ISSN 1073-0516. doi: 10.1145/1067860.1067867. URL https://doi.org/10.1145/1067860.1067867.

Dan Bohus and Alexander Rudnicky. 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, October 2005. URL https://www.aclweb.org/anthology/H05-1029.

Gies Bouwman, Janienke Sturm, and Lou Boves. 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)*, 1999. doi: 10.1109/ICASSP.1999.758170.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proc. Conference on Artificial Intelligence (AAAI)*, 2010. URL http://rtw.ml.cmu.edu/papers/carlson-aaai10.pdf.

Zhiyuan Chen and Bing Liu. *Lifelong Machine Learning, Second Edition*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2018. doi: 10.2200/S00832ED1V01Y201802AIM037. URL https://doi.org/10.2200/S00832ED1V01Y201802AIM037.

Fred D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340, 1989. ISSN 02767783. URL http://www.jstor.org/stable/249008.

Marie-Catherine de Marneffe, Scott Grimm, and Christopher Potts. Not a simple yes or no: Uncertainty in indirect answers. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 136–143, 2009. ISBN 978-1-932432-64-0.

Alexiei Dingli and Darren Scerri. Building a hybrid: Chatterbot – dialog system. In *Proc. International Conference on Text, Speech, and Dialogue (TSD)*, pages 145–152, 2013. ISBN 978-3-642-40585-3.

Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A. Smith, and Mari Ostendorf. Sounding board: A user-centric and content-driven social chatbot. In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)*, pages 96–100, June 2018. doi: 10.18653/v1/N18-5020. URL https://www.aclweb.org/anthology/N18-5020.

Sarah E. Finch, James D. Finch, Ali Ahmadvand, Ingyu Choi, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, and Jinho D. Choi. Emora: An inquisitive social chatbot who cares for you. In *Alexa Prize Proceedings*, 2020.

Ajda Gokcen and Marie-Catherine de Marneffe. I do not disagree: leveraging monolingual alignment to detect disagreement in dialogue. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 94–99, 2015.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11: 10–18, November 2009. doi: http://doi.acm.org/10.1145/1656274.1656278.

Jahn Heymann, Oliver Walter, Reinhold Haeb-Umbach, and Bhiksha Raj. Iterative Bayesian word segmentation for unsupervised vocabulary discovery from phoneme lattices. In *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, pages 4057–4061, 2014.

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 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, 2010. ISBN 9781932432855. URL https://www.aclweb.org/anthology/W10-4304.

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

Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)*, pages 851–861, 2015. doi: 10.3115/v1/N15-1086. URL http://aclweb.org/anthology/N15-1086.

Hartwig Holzapfel, Daniel Neubig, and Alex Waibel. A dialogue approach to learning object descriptions and semantic categories. *Robotics and Autonomous Systems*, 56(11):1004–1013, 2008.

Yukiko Kageyama, Yuya Chiba, Takashi Nose, and Akinori Ito. Improving user impression in spoken dialog system with gradual speech form control. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 235–240, 2018. doi: 10.18653/v1/W18-5026. URL https://aclanthology.org/W18-5026.

Takahiro Kobori, Mikio Nakano, and Tomoaki Nakamura. 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, 2016. doi: 10.18653/v1/w16-3646. URL http://www.aclweb.org/anthology/W16-3646.

Kazunori Komatani and Tatsuya Kawahara. 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, 2000. doi: 10.3115/990820.990888. URL `https://doi.org/10.3115/990820.990888`.

Kazunori Komatani and Mikio Nakano. User impressions of questions to acquire lexical knowledge. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 147–156, 1st virtual meeting, July 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.sigdial-1.19`.

Kazunori Komatani, Tsugumi Otsuka, Satoshi Sato, and Mikio Nakano. Question selection based on expected utility to acquire information through dialogue. In *Proc. International Workshop on Spoken Dialogue System Technology (IWSDS)*, pages 27–38, 2016. doi: 10.1007/978-981-10-2585-3_6.

Kazunori Komatani, Yuma Fujioka, Keisuke Nakashima, Katsuhiko Hayashi, and Mikio Nakano. Knowledge graph completion-based question selection for acquiring domain knowledge through dialogues. In *Proc. International Conference on Intelligent User Interfaces (IUI)*, pages 531–541, 2021. ISBN 9781450380171. doi: 10.1145/3397481.3450653. URL `https://doi.org/10.1145/3397481.3450653`.

Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, and Gary Geunbae Lee. Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication*, 51(5):466 – 484, 2009.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Learning through dialogue interactions by asking questions. In *Proc. International Conference on Learning Representations (ICLR)*, 2017. URL `https://openreview.net/pdf?id=rkE8pVcle`.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 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, 2019. doi: 10.18653/v1/D19-1187. URL `https://doi.org/10.18653/v1/D19-1187`.

Gale M. Lucas, Jill Boberg, David Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. Getting to know each other: The role of social dialogue in recovery from errors in social robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 344–351, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349536. doi: 10.1145/3171221.3171258. URL `https://doi.org/10.1145/3171221.3171258`.

Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. 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, September 2019. URL `https://www.aclweb.org/anthology/W19-5903`.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 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, 2012. doi: 10.21437/Interspeech.2012-73.

Helen Meng, P. C. Ching, Shuk Fong Chan, Yee Fong Wong, and Cheong Chat Chan. ISIS: An adaptive, trilingual conversational system with interleaving interaction and delegation dialogs. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 11(3):268–299, 2004.

Gregoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539, March 2015. ISSN 2329-9290.

Tom M. Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao nad Kathryn Mazaitis, Thahir Mohamed, Ndapa Nakashole, Emmanouil Antonios Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greavesand, and Joel Welling. Never-ending learning. In *Proc. Conference on Artificial Intelligence (AAAI)*, 2015. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10049`.

Mikio Nakano and Kazunori Komatani. A framework for building closed-domain chat dialogue systems. *Knowledge-Based Systems*, 204:106212, 2020. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2020.106212. URL `http://www.sciencedirect.com/science/article/pii/S0950705120304287`.

Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. Toward lexical acquisition during dialogues through implicit confirmation for closed-domain chatbots. In *Proc. of Second Workshop on Chatbots and Conversational Agent Technologies (WOCHAT)*, 2016. URL `http://workshop.colips.org/wochat/@iva2016/documents/RP-272.pdf`.

Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. 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, 2017. doi: 10.18653/v1/W17-5507.

Tsugumi Otsuka, Kazunori Komatani, Satoshi Sato, and Mikio Nakano. 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, August 2013. URL `http://www.aclweb.org/anthology/W/W13/W13-4009`.

Ioannis Papaioannou and Oliver Lemon. Combining chat and task-based multimodal dialogue for more engaging HRI: A scalable method using reinforcement learning. In *Proc. ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 365–366, 2017. ISBN 978-1-4503-4885-0.

Aasish Pappu and Alexander Rudnicky. 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, June 2014. doi: 10.3115/v1/W14-4326. URL `https://www.aclweb.org/anthology/W14-4326`.

Cathy Pearl. *Designing Voice User Interfaces: Principles of Conversational Experiences*. O'Reilly Media, Inc., 1st edition, 2016. ISBN 1491955414.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proc. European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–325, 2021. URL `https://aclanthology.org/2021.eacl-main.24`.

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

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2021–2030, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.183. URL `https://www.aclweb.org/anthology/2020.acl-main.183`.

Janienke Sturm, Els den Os, and Lou Boves. 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, 1999.

Ming Sun, Yun-Nung Chen, and Alexander I. Rudnicky. Learning OOV through semantic relatedness in spoken dialog systems. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1453–1457, 2015. doi: 10.21437/Interspeech.2015-347.

Yasuhiro Takahashi, Kohji Dohsaka, and Kiyoaki Aikawa. An efficient dialogue control method using decision tree-based estimation of out-of-vocabulary word attributes. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 813–816, 2002.

Ryu Takeda and Kazunori Komatani. Attribute prediction of unknown lexical entities based on mixture of Bayesian segmentation model. In *Proc. of Life Long Learning for Spoken Language Systems Workshop*, 2019.

Ryu Takeda, Kazunori Komatani, and Alexander I. Rudnicky. Word segmentation from phoneme sequences based on Pitman-Yor semi-Markov model exploiting subword information. In *Proc. of IEEE Workshop on Spoken Language Technology (SLT)*, pages 763–770, 2018.

Stefan Ultes. 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, September 2019. URL `https://www.aclweb.org/anthology/W19-5902`.

Stefan Ultes and Wolfgang Minker. 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, 2014. doi: 10.3115/v1/W14-4328. URL `https://aclanthology.org/W14-4328`.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 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, 1997. doi: 10.3115/976909.979652. URL https://aclanthology.org/P97-1035.pdf.

Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. Knowledge base completion via search-based question answering. In *Proc. International Conference on World Wide Web (WWW)*, pages 515–526, 2014. ISBN 978-1-4503-2744-2. doi: 10.1145/2566486.2568032. URL http://doi.acm.org/10.1145/2566486.2568032.

Jason Weston. Dialog-based language learning. In *Proc. International Conference on Neural Information Processing Systems (NIPS)*, pages 829–837, 2016. ISBN 978-1-5108-3881-9. URL http://dl.acm.org/citation.cfm?id=3157096.3157189.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 3351–3357, 2017. URL https://dl.acm.org/doi/10.5555/3298023.3298055.

Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. Augmenting end-to-end dialog systems with commonsense knowledge. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 4970–4977, 09 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16573.

Dian Yu, Michelle Cohn, Yi Mang Yang, Chun Yen Chen, Weiming Wen, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Sam Davidson, Ashwin Bhandare, and Zhou Yu. Gunrock: A social bot for complex and engaging long conversations. In *Proc. Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 79–84, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3014. URL https://www.aclweb.org/anthology/D19-3014.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. Strategy and policy learning for non-task-oriented conversational systems. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 404–412, September 2016.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4623–4629, 7 2018. doi: 10.24963/ijcai.2018/643. URL https://doi.org/10.24963/ijcai.2018/643.