Asking Follow-Up Clarifications to Resolve Ambiguities in Human-Robot Conversation

Fethiye Irmak Doğan, Ilaria Torre and Iolanda Leite
KTH Royal Institute of Technology, Stockholm, Sweden
fidogan@kth.se; ilariat@kth.se; iolanda@kth.se

***Abstract*—When a robot aims to comprehend its human partner's request by identifying the referenced objects in Human-Robot Conversation, ambiguities can occur because the environment might contain many similar objects or the objects described in the request might be unknown to the robot. In the case of ambiguities, most of the systems ask users to repeat their request, which assumes that the robot is familiar with all of the objects in the environment. This assumption might lead to task failure, especially in complex real-world environments. In this paper, we address this challenge by presenting an interactive system that asks for follow-up clarifications to disambiguate the described objects using the pieces of information that the robot could understand from the request and the objects in the environment that are known to the robot. To evaluate our system while disambiguating the referenced objects, we conducted a user study with 63 participants. We analyzed the interactions when the robot asked for clarifications and when it asked users to re-describe the same object. Our results show that generating follow-up clarification questions helped the robot correctly identify the described objects with fewer attempts (i.e., conversational turns). Also, when people were asked clarification questions, they perceived the task as easier, and they evaluated the task understanding and competence of the robot as higher. Our code and anonymized dataset are publicly available[1].**

***Index Terms*—Resolving Ambiguities, Follow-Up Clarifications, Referring Expressions**

I. Introduction

In Human-Robot Conversation, verbal ambiguities or comprehension failures are inevitable. Ambiguities might happen because of unclear user instructions, speech recognition errors, or simply because the environment is so complex that providing a clear request is difficult for the user. For instance, when a user describes an object and asks a robot to hand it over, the described object might be unknown to the robot, the object description might fit more than one object, and/or the spatial configuration of the environment might be very similar for many objects. When facing these uncertainties, most of the existing language-based systems (such as Amazon Alexa or Siri) simply ask users to repeat the request over and over again. In the previous example, given the fact that people prefer to use clarifications rather than just repeating or rephrasing requests if there are ambiguities [1], asking the users to re-describe the object could potentially increase users' frustration and the likelihood of task failure in cases where rephrasing is difficult.

In the particular case of describing objects, people tend to use referring expressions (i.e., an object description that specifies the object's distinguishing features). To identify the objects in a user instruction, the robot needs to comprehend the referring expressions. Existing approaches for comprehending referring expressions typically assume that the candidate objects in the environment are given [2] or that these candidates can be found by existing object detection or localization methods [3]–[5]. The robots then select the target object from these candidates. However, it is unrealistic to assume that the robot that must comprehend a referring expression is familiar with all of the object names or that the objects in the environment can be detected or localized by state-of-the-art object detectors, especially in real-world scenarios. Therefore, this assumption can cause a failure for previous approaches when the robot encounters unknown objects in the real world. A possible solution in these cases would be for the robot to utilize the information that it was able to understand from the user request and ask for follow-up clarifications using known objects (e.g., detectable ones) and concepts (e.g., color or shape information) in the environment.

In this paper, we propose a system that attempts to clarify the ambiguities in users' referring expressions. When there are ambiguities because of similar or unknown objects in the environment, our system uses the parts of the user request that the robot could understand to form a hypothesis about which object the user intends to describe. More specifically, the robot first uses the understood information to find the objects that might fit the request. For instance, in Figure 1, when the user asks to pick up the green vegetable, the robot might not be familiar with the vegetable concept, but it might have encountered green objects before. In this case, our system enables the robot to identify the green objects and then generate clarification questions using the other detectable objects in the environment and their spatial relations (e.g., "Is the green vegetable to the left of the knife?"). These clarification questions enable the robot to identify the detectable (e.g., spoon, apple, banana) and undetectable (e.g., artichoke, basil, lemon) objects that might fit the description. To our knowledge, our system is the first one aiming to disambiguate

This work was partially funded by grants from the Swedish Research Council (2017-05189), the Swedish Foundation for Strategic Research (SSF FFL18-0199), the S-FACTOR project from NordForsk, the Digital Futures Research Center, the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH, and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. We are grateful to Liz Carter, Ignacio Torroba, Ioanna Mitsioni, and Sarah Gillet for their valuable feedback and Rasmus Rudling for his contribution to the experiment.

[1]https://github.com/IrmakDogan/Resolving-Ambiguities

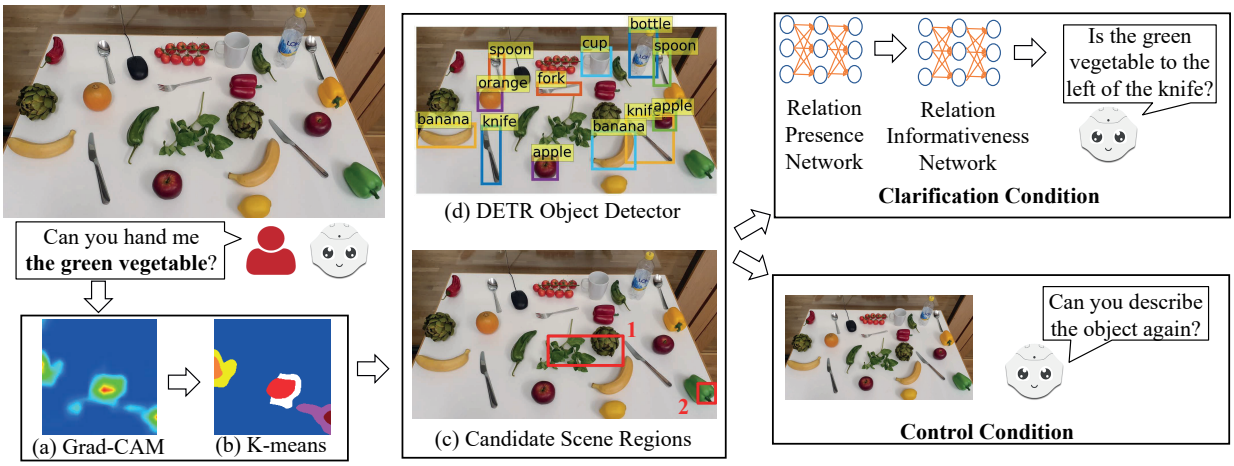


Fig. 1. Overview of our system to find the described objects and to generate clarification questions for resolving ambiguities.

referred objects with clarification questions using the parts of the request the robot could understand and the objects in the environment known by the robot.

In our approach, we first identify the regions that are specified in the user expression using our previous method leveraging explainability [6]. Then, if the user's description can potentially match objects in more than one region, we suggest generating clarification questions for each of these potential target regions instead of asking the user to re-describe the object. To generate the follow-up clarifications, we employ another prior method of ours [7] (i.e., generating unambiguous spatial referring expressions) and describe each potential target region with their spatial relations to other objects (detected by an object detector). To evaluate the impact of asking for follow-up clarifications on the interaction, we conducted a user study with 63 participants. The study results show that when the robot aimed to disambiguate the target objects with clarification questions instead of requesting another description of the same object, it could find the described object more often with fewer attempts (i.e., fewer conversational turns). Further, when the robot asked for follow-up clarifications, people perceived the task as easier, and they evaluated the robot's task understanding and competence as higher.

A. Related Work

Referring expressions have been of interest for many robotics applications both for comprehending users' expressions [2]–[6], [8]–[12] and generating unambiguous object descriptions [7], [8], [13]–[19].

To comprehend referring expressions, it has been suggested that the interactive clarification process can help resolve ambiguities and improve task accuracy [3], [4]. To disambiguate the described objects, Hatori et al. [3] proposed a method for computing the similarity between the embedding of the input instructions and each candidate object. When there were uncertainties, the suggested system visualized the most probable target objects to the users and asked *"Which one?"* to identify the referred object. In another work, Shridhar et al. [4], [5] proposed a method to comprehend referring expressions using Long Short-Term Memory (LSTM) Networks [20]. Similar to our approach, the system generated disambiguation questions while identifying the referred objects. However, in this work, the candidate objects were still limited to the ones suggested by the DenseCap object localization module [21]. On the other hand, our work aims to resolve ambiguities using the information that the robot could understand from the users' requests (with the attention map of the explainability module without any candidate object constraints) and asking for clarifications using the known objects of the robot. The work of Shridhar et al. [4], [5] compared the impacts of asking different types of clarification questions (i.e., the robot pointed at the object by asking whether it was the described object or asked object-specific questions) with 24 participants. However, our focus is to analyze the differences between when the robot asks for clarifications and when it asks users to re-describe the object. Asking follow-up questions has been further studied for dialog navigation [22] and clarifying different entities of requests (e.g., recipient, room, etc.) [23].

While asking for follow-up clarifications, generating referring expressions is critical to disambiguate each potential target region. To uniquely describe the target objects, previous studies on referring expression generation commonly utilized spatial relations [13], [14], [17], [18]. For instance, Kunze et al. [17] suggested an algorithm to generate different types of spatial referring expressions (e.g., Set-Relative, Proximal, Distal, etc.) and trained a classifier to determine which algorithm the robot should use while interacting with users. Our previous work [7] suggested a two-stage learning based approach for spatially describing objects in a natural and unambiguous manner. In this paper, we employ our previous spatial referencing method in the clarification questions to describe the potential target regions (details in Section II-C).

B. Contributions

- To endow robots with the ability to disambiguate referred objects in the user's request, we suggest a system that

asks for follow-up clarifications. To our knowledge, our system is the first one that aims to revolve ambiguities by utilizing the parts of the users' request understood by the robot and interactively asking for clarifications using the objects known by the robot in the environment.

- To evaluate the impacts of asking for follow-up clarifications, we conducted a study with 63 participants. We analyzed the interactions by comparing differences between when the robot asked for clarifications and when it asked for another description of the same object in cases of ambiguities. With this comparison, we aim to see whether generated follow-up clarifications can be helpful at all to resolve ambiguities, which to our knowledge has not been investigated before.

II. Asking for Clarifications to Find the Described Object

Given a scene and an object description, our goal is to generate follow-up clarification questions when the robot is uncertain about the described object. To achieve this, we first find the candidate regions in the scene described in the expression, following the approach introduced in Section II-A. If the user's object description matches with a single candidate region (i.e., unambiguous request), we suggest this region as the target object region (see Section II-B). However, if there are uncertainties (an ambiguous request, e.g., there are multiple candidate regions fitting to the object description), we generate a clarification question for each candidate using the spatial relations between the candidate regions and the detected objects (see Section II-C for details).

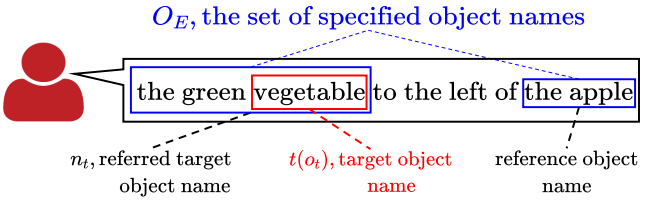
A. Finding the Described Scene Regions

Our previous work [6] proposed a method leveraging explainability to find the regions containing the described object in the scene. Given an RGB image of a scene and a user referring expression r describing an object in the scene, it uses the off-the-shelf image captioning module of Grad-CAM [24] explainability method to generate a heatmap H of the image. This heatmap visually highlights the areas that contribute to the output predictions for the given caption. In the task of finding the described target object, the captions correspond to the user referring expressions – see the generated heatmap for a given scene and a referring expression in Figure 1(a).

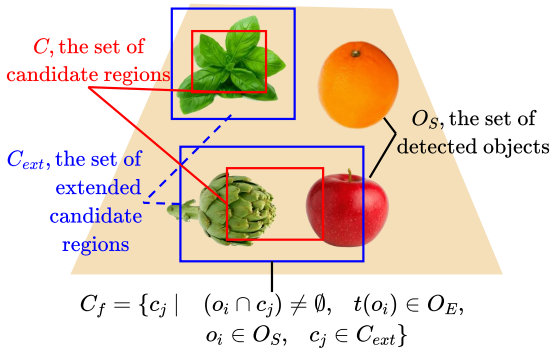
After obtaining the heatmap H, the number of unconnected areas n in H is determined using the 2D connectivity of the pixels. Then, to separate the active regions in H, which might have some overlapping active regions, K-means clustering is applied to H with n clusters – see Figure 1(b). Finally, the set of candidate regions (denoted as C) that correspond to the described objects is chosen as the set of two bounding boxes covering the two most active clusters – the red boxes shown in Figure 1(c) and Figure 2(b) correspond to C in these examples.

B. Identifying the Target Object

To determine which of the suggested candidate regions contains the target object, we first identify the set of detectable



(a) For a given example expression, the corresponding items for O_E, n_t, $t(o_t)$, and the reference object.



(b) For the given example expression in 2(a), an example illustrating the corresponding items for C, C_{ext}, C_f and O_S.

Fig. 2. For a given expression, an example illustrating the defined notations.

objects O_S in the scene using the off-the-shelf DEtection TRansformer (DETR) Object Detector [25] (the detected objects shown in Figure 1(d) correspond to O_S in this example). Then, we parse the user referring expression r with the spaCy natural language processing library [26] to obtain the set of specified object names O_E in the expression. For example, when the referring expression r is 'the green vegetable to the left of the apple', O_E would be {'the green vegetable', 'the apple'}. A schematic overview of O_S and O_E is also depicted in Figure 2.

After finding O_S and O_E, we extend each candidate box in C to capture the reference object names within r that are used as references for the target's location (e.g., if r is 'the green vegetable to the left of the apple', 'the apple' is called as the reference object). We experimentally observe that extending each candidate region by half of its width/height is enough to contain potential reference objects. Therefore, each candidate in C is extended horizontally by half of its width and vertically by half of its height, and the set of extended candidates is denoted as C_{ext}.

After obtaining (i) the set of detectable objects O_S, (ii) the one containing the object names from the referring expression O_E, and (iii) C_{ext} denoting the extended candidate regions, we define another set C_f (the examples to these notations shown in Figure 2). This set gathers the extended candidate regions from C_{ext}, $c_j \in C_{ext}$, that contain at least one object (partially or fully) from O_S whose name is also a member of O_E:

$$C_f = \{c_j \mid (o_i \cap c_j) \neq \emptyset, \quad t(o_i) \in O_E, \quad o_i \in O_S, \quad c_j \in C_{ext}\}, \tag{1}$$

where $t(o_i)$ represents the object name of o_i, and $(o_i \cap c_j) \neq \emptyset$ imposes the required intersection between o_i and c_j. Once C_f has been obtained, there are two possible scenarios that require different approaches to identify the target object:

- If C_f is empty or contains multiple candidate regions (**ambiguous request**), we suggest that the existing systems could not identify one unique target object region. We show how we deal with this case in Section II-C.
- If C_f contains only one candidate box c_t (**unambiguous request**), this region is then proposed as the one containing the target object. The remainder of this section explains how to identify the target object o_t from c_t in this case.

To obtain the target object o_t contained in c_t, we first identify its name by assuming that the beginning of the object description coincides with the target's name. The first member of O_E is then used as the object's referred name (n_t), and the name of the target object $t(o_t)$ is suggested as the noun in n_t. For instance, when the referring expression r is 'the green vegetable to the left of the apple' and O_E is {'the green vegetable', 'the apple'}, n_t and $t(o_t)$ would be 'the green vegetable' and 'vegetable' respectively – see Figure 2(a).

If $t(o_t) \notin \{t(o_i) \mid o_i \in O_S\}$ (**undetectable target object** such as an artichoke, which can not be localized with existing object detectors), we use another off-the-shelf transformer-based network, MDETR [27], to obtain o_t from c_t. This network takes the target object region c_t and the input expression r and matches the given descriptions with the corresponding areas in the scene without requiring a fixed vocabulary. After matching the regions of c_t with the given expression r, we suggest the region matched with n_t as the target object o_t.

On the other hand, if $t(o_t) \in \{t(o_i) \mid o_i \in O_S\}$ (**detectable target object**, such as an apple or a banana which can be localized by the DETR object detector), we determine the target o_t as the object from O_S which is closest to c_t based on their Euclidean distances to the centroid of the region, and has the name $t(o_t)$:

$$o_t \leftarrow \underset{o_i \in O_S \ \wedge \ t(o_i)=t(o_t)}{\arg\min} ||o_i - c_t||, \tag{2}$$

where $||o_i - c_t||$ shows the Euclidean distance between the center of masses of object o_i and the target candidate box c_t.

After obtaining o_t either from MDETR or from the closest object in O_S, we propose o_t as the target object specified in the expression r.

C. Asking for Follow-Up Clarifications

When the system cannot identify the target object region with the approach from the previous section (i.e., an ambiguous request, C_f is either empty or contains more than one region), it generates follow-up clarification questions to disambiguate the potential regions.

To generate follow-up clarifications, we first define two new sets, C_{clf} and O_{all}. C_{clf} represents the probable regions matching with the expression r, and O_{all} denotes the objects that can be used to generate the clarification question considering their spatial arrangement.

If $t(o_t) \notin \{t(o_i) \mid o_i \in O_S\}$ (**undetectable target object**), we set $C_{clf} = C$. In this case, each candidate region in C is assumed to be an object with object name $t(o_t)$. Then, the combined set O_{all} is formed by the union of these objects and the already detected ones in O_S:

$$O_{all} = O_S \cup \{o_j \leftarrow c_j \mid \quad \forall c_j \in C\}. \tag{3}$$

On the other hand, if $t(o_t) \in \{t(o_i) \mid o_i \in O_S\}$ (**detectable target object**), we set $O_{all} = O_S$. Then, we define C_{clf} as the set of objects in O_S that are closest to the candidates in C and have object names $t(o_t)$:

$$C_{clf} = \left\{ \underset{o_i \in O_S \ \wedge \ t(o_i)=n_t}{\arg\min} ||o_i - c_j||, \quad \forall c_j \in C \right\}. \tag{4}$$

After obtaining C_{clf} and O_{all}, we generate a clarification question for each possible region employing our previous method [7] that aims to describe objects unambiguously with spatial referring expressions.

Firstly, the spatial relations among the objects in O_{all} are detected using the Relation Presence Network (RPN), trained as in our prior work [7]. This network takes pairs of objects as input and outputs the spatial relations between them. In our prior work, these relations could belong to six different classes: 'to the left', 'to the right', 'in front', 'behind', 'on top', 'at the bottom'. In this work, we discard the 'on top' and 'at the bottom' relations and include 'close to'.

After finding all of the spatial relations among the objects in O_{all}, the suggested relations are provided to the Relation Informativeness Network (RIN), trained as in its suggested paper [7]. This network takes pairs of objects and their spatial relationship (obtained from RPN) and calculates a confidence value showing the informativeness of this spatial relation, i.e., how much the spatial relation uniquely describes the object.

Using the confidence value obtained from RIN, we find the most informative relation s and reference object o_r that unambiguously describe each possible region in C_{clf}. Then, we form a yes/no question Q by concatenating the referred target object name n_t, the most informative spatial relation s and the name of the reference object o_r:

$$Q \leftarrow n_t \oplus s \oplus t(o_r), \quad o_r \in O_{all} \tag{5}$$

where $\oplus$ represents the concatenation of the object and relation names. If the system receives an affirmative answer to the yes/no question Q for a region in C_{clf}, this region is suggested as the target object region c_t, and the target object o_t is obtained from c_t as explained in the previous section.

The suggested system to generate follow-up clarifications is presented as **Clarification Condition** in Figure 1 – see Algorithm 1 in the Appendix for the overall procedure summary.

D. Asking the User to Describe the Target Object Again

As an alternative solution, when the suggested approach can not identify the target object region c_t (i.e., an ambiguous

request; C_f is either empty or contains more than one candidate), another description of the same target object can be provided to the proposed system instead of asking follow-up clarifications. In this case, the procedure explained in Section II-A and II-B can be iteratively applied until C_f contains only one candidate, and this candidate can be suggested as the target object region. This alternative approach is shown as the **Control Condition** in Figure 1 – see Algorithm 2 in the Appendix for the summary of the procedure.

III. Experiment

To evaluate the impact of asking for follow-up clarification questions as a way to facilitate the interaction, we designed a task with a lot of ambiguities (many similar objects and similar spatial relationships between them) where users had to describe objects to a robot. When there were ambiguities in the object descriptions, the robot either asked clarification questions (**Clarification Condition** explained in Section II-C) or asked users to describe the object again (**Control Condition** described in Section II-D). In this study, we considered that generating clarification requests has been suggested as necessary for robots to resolve referential ambiguities [28], and we formulated the following hypotheses:

H1. When ambiguities occurred, we expected that asking a clarification question would result in the robot successfully identifying the described objects more often than when it asked for another object description.

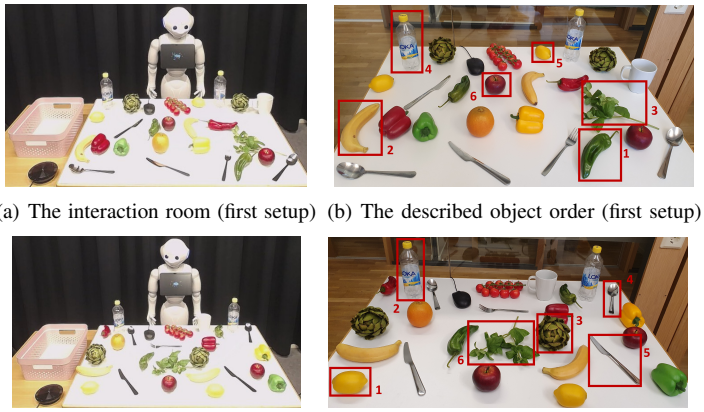
H2. We hypothesized that when the robot used clarification questions in an attempt to resolve ambiguities, the number of attempts (i.e., conversational turns) required to identify the described objects would be fewer than when the robot simply asked the user to re-describe the same object again.

H3. Because people use clarifications instead of repeating or rephrasing the expressions in task-oriented dialogues [1], we hypothesized that people would perceive the task as easier when the robot asked for clarification instead of another description of the same object to resolve uncertainties.

H4. Because clarification questions are described as critical for ensuring mutual understanding in dialog systems [29], we expected that when the robot asked for follow-up clarifications to resolve ambiguities, people would think the robot comprehended their instructions better, and they would perceive the competence of the robot as higher.

A. Study Design

To evaluate the effects of asking follow-up clarifications, we designed two different ambiguous table setups (i.e., with many similar objects and similar spatial relations on a table). The order of these setups was the same for all participants. In other words, they always started the experiment with the first setup shown in Figure 3(a) and continued with the second setup presented in Figure 3(c), but the condition in which they started the experiment (either Clarification or Control) changed between participants. In each setup, participants were asked to uniquely describe six target objects to the robot. The first and second setups contained exactly the same objects with



(a) The interaction room (first setup) (b) The described object order (first setup)

(c) The interaction room (second setup) (d) The described object order (second setup)

Fig. 3. The interaction room with the first and second table setup in 3(a) and 3(c), and the order of the objects described during the study in 3(b) and 3(d).

different spatial arrangements and different target objects to be described. Each setup was balanced with three detectable (i.e., the object names could be identified by an object detector) and three undetectable target objects. The target object order in each setup was provided to the participants for a fair evaluation (see Figure 3(b) and 3(d)). For instance, all of the participants started the experiment by describing the first object in the red box in Figure 3(b). After each object, the participants took the described object from the table and put it in a basket.

We used a within-subject design. We counterbalanced in which conditions that participants started the experiment such that they either started with the Clarification Condition and continued with the Control Condition or vice versa.

Clarification Condition: When there were ambiguities in the user's description, the robot generated clarification questions to identify the target object as explained in Section II-C – see the top right corner of Figure 1. After the initial participant description of an object, the robot considered the first two candidate regions, and it could ask at most two follow-up questions to identify each target object. The robot stopped asking for clarifications after the second time to avoid frustrating participants.

Control Condition: When the robot was uncertain about the described object (i.e., ambiguous request; either candidate list C_f is empty or contains more than one regions matching with the description), it asked participants to describe the object again as presented in Section II-D – see the bottom right corner of Figure 1. After the user described an object for the first time, the robot could ask users to describe each target object again up to two more times. If it could not identify the target object after two re-descriptions, it asked participants to move on the next object to avoid frustration.

In our experimental design, we strived to have a fair comparison between the control and clarification conditions. In both conditions, we extended our previous method [6] (shown to outperform MattNet [30] for challenging undetectable objects) with DETR and MDETR to better identify objects and utilize

the detectable ones, as explained in Section II.

B. Procedure

The experiment was conducted in two different rooms (the interaction room shown in Figure 3(a) and 3(c) and the questionnaire room where the participants filled out the consent forms and questionnaires) located at the campus of KTH Royal Institute of Technology and took around 30 to 40 minutes for each participant. All experiments were conducted by the same experimenter (the first author) with the same humanoid Pepper robot[2].

First, participants signed a consent form and answered demographic questions. Then, in the interaction room, participants were informed that the experiment would consist of two parts, and in each part, the Pepper robot would try to find the described objects with different behaviors. They were instructed to uniquely describe objects from their viewpoints in the provided object order and asked to stand across from the Pepper robot on the other side of the table. The participants were told that when the robot thought it could find the described object, it would show a region in a red box on its tablet, and they should tell the robot whether the suggested object was correct or incorrect. They were also informed that they should take each described object from the table and put it in a basket. After the instructions, the experimenter left the interaction room and monitored the interaction from another room with an external camera.

When the participants finished the first part of the experiment, they completed a questionnaire about their interaction and then continued the experiment with the second table arrangement, following the same instructions. After describing the objects in the second setup, the participants filled out another questionnaire regarding only the second part of the interaction. After the second questionnaire and debriefing about the experiment, the participants received compensation for their time.

During the interaction, we used the Pepper robot's built-in text-to-speech and basic awareness components. To capture the participants' object descriptions, we used the Google speech recognition engine [31]. When the engine failed to capture the object descriptions correctly, the experimenter corrected the given expressions. Therefore, according to Beer et al. [32], the autonomy level was '*Shared Control with Human Initiative*' to capture the object descriptions, '*Full Automation*' to generate clarification questions and to find the described objects and '*Decision Support*' for picking up the described objects, capturing the replies to the clarification questions and getting the responses for the proposed target objects. For the parts that required '*Shared Control with Human Initiative*' and '*Decision Support*', the experimenter was trained in a pilot study with seven users.

C. Measures

To evaluate our hypotheses, we collected the objective measures during the interaction and the subjective measures with questionnaires after the participants interacted with the robot in each condition.

1) Objective Measures: To assess **H1** (i.e., finding an ambiguous target object more often with clarification questions) and evaluate task accuracy, we counted the total number of times that the robot made correct and incorrect predictions. The cases where the robot could not propose an object after two attempts at resolving ambiguities (by asking two clarification questions or asking the users twice to describe the object again) were counted as an unknown object.

In order to evaluate the second hypothesis **H2**, we collected the number of attempts required by the robot to find the described object in each condition (by asking for either clarification questions or another description of the same object).

2) Subjective Measures: To answer **H3** and evaluate the perceived task difficulty, the participants evaluated the following statement on a seven-point response item: '*The task was easy to perform*'.

To investigate the last hypothesis **H4**, we evaluated perceptions of the robot. In the post-experiment questionnaire, we asked participants to rate the following statement on a seven-point scale: '*The robot could understand my instructions*'. Moreover, to assess the competence of the robot perceived by participants, we asked participants to evaluate the items in the Competence factor of The Robotic Social Attribute Scale (RoSAS) [33] on a seven-point scale.

D. Participants

The experiment was advertised through different channels (e.g., posters in universities, invitation links on social media), and 69 participants joined the experiment. Out of the 69 participants, three participants could not complete the experiment because of the robot's hardware problems, and three participants did not comply with the instructions. These six interactions were excluded from the analysis. Among the remaining 63 participants from 24 different countries, 23 participants identified themselves as female, and 40 of them identified as male. Most of the participants reported a bachelor's degree. They ranged in age from 18 to 49 years (M: 26.76, STD: 5.85). Most of the participants had interacted with a robot before the experiment but were not using one in their everyday lives (e.g., at home/work).

During the study, the conditions that participants started the interaction were counterbalanced. Therefore, 32 participants started the experiment with the clarification condition, and 31 started with the control condition.

IV. Results

A. Objective Results

1) Task Accuracy: Among the 378 trials (63 participants, 6 objects per condition) in the clarification condition, the robot could find the described objects correctly in 292 trials, incorrectly identified the objects in 58 trials, and could not suggest any object after two follow-up questions in 28 trials. In contrast, in the control condition, the robot's predictions were correct in 164 trials and incorrect in 56 trials; in 158

[2]https://www.softbankrobotics.com/emea/en/pepper

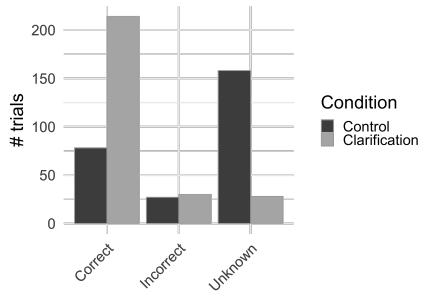


Fig. 4. Number of correct, incorrect and unknown trials in each condition.

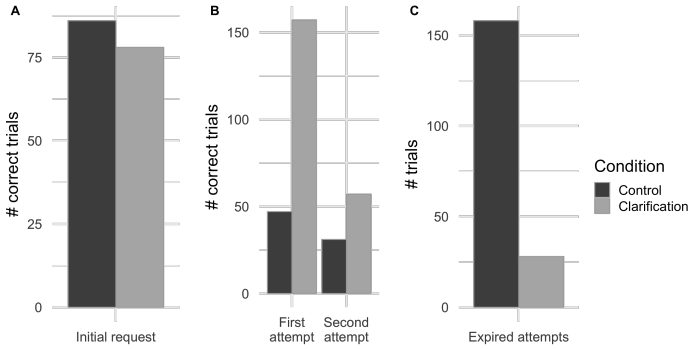


Fig. 5. (A) The number of correct trials after the user's initial object descriptions without clarification or re-description. (B) The number of trials for each condition with correct object identifications after the first and second disambiguation attempts. (C) The number of trials for each condition in which the robot ran out of attempts without proposing an object.

trials, the robot could not manage to suggest any object after two requests for the user to describe the object again.

To analyze whether asking a clarification question in the case of ambiguities led to higher object identification accuracy (**H1**), we conducted a 2-sample chi-square test of independence comparing the outcomes in the clarification and the control conditions. After eliminating all of the trials without ambiguities (i.e., all of the trials for which the object was identified by the robot without the need for a clarification question or another object description), we found that the test was statistically significant ($\chi^2(2)$ = 154.25, $p < .001$). As can be seen from Figure 4, there were more correct trials in the clarification condition and more unknown trials (i.e., trials where the robot did not manage to suggest any object after 2 attempts of resolving ambiguities) in the control condition. Thus, we find evidence supporting **H1**.

2) Number of Attempts to Identify the Correct Object: To assess **H2**, we looked at how many attempts it took the robot to identify the correct object in the two conditions. First, we confirmed our expectation that there was no significant difference between the two conditions in the number of cases where the robot correctly identified the target object without a need for a clarification or another object description ($\chi^2(1) = 0.39, p = .53$), as shown in Figure 5(A). Next, we found that there were many more correct identifications after the first attempt of clarification in the clarification condition relative to the control condition ($\chi^2(1) = 63.34 p < .001$) (Figure 5(B)). Finally, we found the robot ran out of attempts to identify the correct object many more times in the control condition than in the clarification condition ($\chi^2(1) = 90.86, p < .001$) (Figure 5(C)).

3) The Similarity of the Expressions while Re-describing: We measured the similarity of participants' referring expressions when they were asked to describe the same object again in the control condition. To measure the similarity of two expressions while the user describing the same object, we computed an Individual 1-gram BLEU score [34]. Given two expressions, the BLEU score generates a value ranging from 0 (no similarity) to 1 (identical expressions). For instance, when two expressions are *"the knife between the apple and the banana"* and *"the knife next to the apple and the banana"*, the BLEU score is 0.778.

To calculate the BLEU score, we assumed the reference expression was the one obtained when the participant described an object for the first time, and the second or third expressions describing the same object were considered to be the candidate expressions. Then, we computed the BLEU score between the reference expression and each candidate expression. The results suggested a high similarity (M: 0.784, STD: 0.174), suggesting that when participants were asked to describe the same object again, they did not rephrase their expressions much, instead tending to repeat their previous expression.

B. Subjective Results

1) Perception of the Task: To understand whether the perceived difficulty of the task was different in the two conditions (**H3**), we evaluated the item aiming to examine task difficulty in the post-experiment questionnaire. As shown in Figure 6(A), participants in the clarification condition rated the task as easier than in the control condition (M = 5.89 and 4.98 respectively), and an Analysis of Variance revealed that this difference was statistically significant ($F(1)$ = 13.48, $p < .001$).

2) Perception of the Robot: To evaluate the perception of the robot in two conditions (**H4**), we checked the perceived understanding and perceived competence of the robot.

For the robot's perceived understanding, we looked at participants' answers to the item *"The robot could understand my instructions"*. Ratings in the clarification condition were significantly higher than in the control condition (M = 5.79 and 3.86, respectively; $F(1)$ = 65.97, $p < .001$), as shown in Figure 6(B).

While evaluating the perceived competence of the robot, the six relevant items on the RoSAS scale were averaged to obtain one score per participant per condition. As can be seen in Figure 6(C), the average score in the clarification condition was higher than in the control condition (M = 5.58 and 4.15, respectively) and an Analysis of Variance revealed that this difference was statistically significant ($F(1)$ = 64.29, $p < .001$).

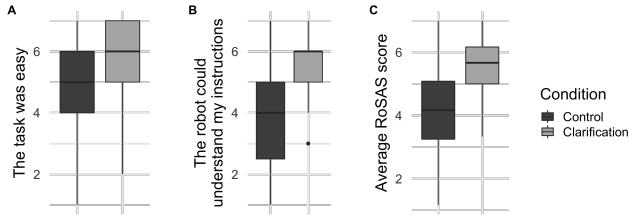


Fig. 6. The subjective results: (A) task difficulty, (B) the robot's perceived understanding, and (C) the RoSAS competence score for the robot.

V. Discussion

Our first hypothesis **H1** predicted that the robot would be able to identify the referred objects more often when it asked for clarifications instead of another description of the same object, and our results supported this hypothesis. As shown in Figure 4, the number of times that the robot predicted the correct objects was significantly more in the clarification condition, and there were fewer trials that resulted in unknown outcomes (i.e., where the robot couldn't propose an object after the second attempt, and the number of maximum attempts to find the target object expired). This shows that when the accuracy of the task is critical in ambiguous environments (e.g., when a medical robot helps a doctor during surgery by handing over the specified surgical instrument when selecting among many similar ones), asking for follow-up clarifications using the parts of the request that were understood and objects known by the robot can help complete the task successfully.

The second hypothesis **H2** expected that the number of attempts required to identify the described objects during ambiguities would be fewer when the robot asked for clarifications, and this hypothesis is supported by the results and shown in Figure 5. First, there were no significant differences between conditions when the robot predicted an object just after the participant's initial object description without a need for clarification or object re-description (Figure 5(A)). This result was expected because two conditions had the same target objects and followed the same procedure when there were no ambiguities (described in Section II-A and II-B). On the other hand, when there was a need for disambiguation, the robot could identify the described object correctly in its first or second follow-up attempts more often in the clarification condition, as shown in Figure 5(B). Moreover, the number of trials with expired attempts where the robot could not propose any objects was significantly higher in the control condition (Figure 5(C)). This could be because when the robot asked users to re-describe the objects in the control condition, the users didn't change their object description much, which is supported by the high BLEU score similarity among the users' expressions while re-describing the objects (presented in Section IV-A3). We observed that some users initially repeated their expressions because they thought the problem was speech-to-text in the control condition, but after the first few requests, they quickly understood that speech-to-text was not the issue. Still, the complexity of the scenes (many similar objects and similar spatial relations) made rephrasing difficult for them and yielded ambiguities. These results suggest that when users are asked to re-describe objects, they might not provide further disambiguating information about the target object, but this information can be obtained with follow-up clarifications, and target objects can be identified with fewer attempts.

In **H3**, it was expected that people would perceive the task as easier when they were asked follow-up clarifications to resolve ambiguities, and the results shown in Figure 6(A) supported this claim. These results are also in line with previous findings in task-oriented dialogues, i.e., people ask clarification questions of each other instead of repeating or rephrasing when there are ambiguities [1].

The last hypothesis **H4** suggested that the robot's task understanding and competence would be perceived as higher when it asked for clarifications instead of re-descriptions to disambiguate the referred objects, and this hypothesis was supported by the results shown in Figure 6(B) and 6(C). These results show that when the perception of the robot is critical for a task with possible uncertainties (e.g., when the robot directs a user with instructions in an ambiguous environment, the user has to perceive the robot as competent to follow the instructions), asking follow-up clarification can be helpful.

One of the main limitations of our system that could be improved in future research is its way of handling user responses to the clarification questions. Specifically, when there were ambiguities, the robot asked yes/no clarification questions to the users, but the responses to these questions could include further information about the target object (in addition to yes/no acknowledgment), especially when the users responded no to the clarification questions. This additional information potentially could be combined with the previous knowledge about the target object to facilitate disambiguation. As an alternative approach, the robot could ask open-ended questions to ensure that users would provide this additional information in all of their responses. We plan to improve our system in this direction in our immediate future work. Another promising direction to improve our system would be to use the collected data to train an end-to-end network and learning to generate follow-up questions with a single-stage approach.

VI. Conclusion

To identify and disambiguate objects described in users' requests, we described a system that asks for follow-up clarifications. Our approach uses the parts of a users' request that could be understood by the robot and the objects known by the robot to generate clarification questions. To evaluate the impacts of asking clarifications to resolve ambiguities, we conducted a user study (N : 63). Our results show that when the robot asked for clarifications to disambiguate the referred objects rather than asking for re-description, it found the target object more often and in fewer attempts (i.e., conversational turns), people evaluated the task as easier, and people perceived the robot's task understanding and competence as higher.

REFERENCES

[1] V. Rieser and J. D. Moore, "Implications for generating clarification requests in task-oriented dialogues," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 2005, pp. 239–246.

[2] A. Magassouba, K. Sugiura, A. T. Quoc, and H. Kawai, "Understanding natural language instructions for fetching daily objects using gan-based multimodal target–source classification," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3884–3891, 2019.

[3] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, "Interactively picking real-world objects with unconstrained spoken language instructions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3774–3781.

[4] M. Shridhar and D. Hsu, "Interactive visual grounding of referring expressions for human-robot interaction," in *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.

[5] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, p. 0278364919897133, 2020.

[6] F. I. Dogan, G. I. Melsion, and I. Leite, "Leveraging explainability for comprehending referring expressions in the real world," *arXiv preprint arXiv:2107.05593*, 2021.

[7] F. I. Doğan, S. Kalkan, and I. Leite, "Learning to generate unambiguous spatial referring expressions for real-world environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4992–4999.

[8] H. Zender, G.-J. M. Kruijff, and I. Kruijff-Korbayová, "Situated resolution and generation of spatial referring expressions for robotic assistants," in *IJCAI*, 2009.

[9] G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes, "Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction," in *Symposium on Language and Robots*, 2007.

[10] T. Kollar, V. Perera, D. Nardi, and M. Veloso, "Learning environmental knowledge from task-based human-robot dialog," in *ICRA*. IEEE, 2013.

[11] R. Paul, J. Arkin, N. Roy, and T. M Howard, "Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators," in *RSS*, 2016.

[12] J. Thomason, A. Padmakumar, J. Sinapov, J. Hart, P. Stone, and R. J. Mooney, "Opportunistic active learning for grounding natural language descriptions," in *Conference on Robot Learning*. PMLR, 2017, pp. 67–76.

[13] R. Moratz and T. Tenbrink, "Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations," *Spatial cognition and computation*, vol. 6, no. 1, pp. 63–107, 2006.

[14] S. Guadarrama, L. Riano, D. Golland, D. Go¨hring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell, "Grounding spatial relations for human-robot interaction," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 1640–1647.

[15] T. Williams and M. Scheutz, "Referring expression generation under uncertainty: Algorithm and evaluation framework," in *INLG*, 2017.

[16] ——, "Referring expression generation under uncertainty in integrated robot architectures," in *RSS Workshop on Human-Centered Robotics: Interaction, Physiological Integration and Autonomy*, 2017.

[17] L. Kunze, T. Williams, N. Hawes, and M. Scheutz, "Spatial referring expression generation for hri: Algorithms and evaluation framework," in *AAAI Fall Symposium on AI and HRI*, 2017.

[18] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock, "Spatial language for human-robot dialogs," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 2, pp. 154–167, May 2004.

[19] A. Magassouba, K. Sugiura, and H. Kawai, "Multimodal attention branch network for perspective-free sentence generation," 2019.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[21] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[22] H. R. Roman, Y. Bisk, J. Thomason, A. Celikyilmaz, and J. Gao, "Rmm: A recursive mental model for dialog navigation," in *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[23] S. Amiri, S. Bajracharya, C. Goktolgal, J. Thomason, and S. Zhang, "Augmenting knowledge through statistical, goal-oriented human-robot dialog," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 744–750.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[26] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.1212303

[27] A. Kamath, M. Singh, Y. LeCun, I. Misra, G. Synnaeve, and N. Carion, "Mdetr–modulated detection for end-to-end multi-modal understanding," *arXiv preprint arXiv:2104.12763*, 2021.

[28] T. Williams and M. Scheutz, "Resolution of referential ambiguity in human-robot dialogue using dempster-shafer theoretic pragmatics." in *Robotics: Science and Systems*, 2017.

[29] M. Gabsdil, "Clarification in spoken dialogue systems," in *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, 2003, pp. 28–35.

[30] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1307–1315.

[31] A. Zhang, "Speech recognition (version 3.8)," 2017. [Online]. Available: https://github.com/Uberi/speech_recognition#readme

[32] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," *Journal of human-robot interaction*, vol. 3, no. 2, p. 74, 2014.

[33] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas): Development and validation," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '17. New York, NY, USA: ACM, 2017, pp. 254–262. [Online]. Available: http://doi.acm.org/10.1145/2909824.3020208

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.