Intelligent Service Robotics (2022) 15:411–426
https://doi.org/10.1007/s11370-022-00411-z

ORIGINAL RESEARCH PAPER

An in-the-wild study to find type of questions people ask to a social robot providing question-answering service

Syed Ali Raza[1] · Jonathan Vitale[1] · Meg Tonkin[1] · Benjamin Johnston[1] · Richard Billingsley[1] · Sarita Herse[1] · Mary-Anne Williams[1]

Received: 7 November 2020 / Accepted: 30 December 2021 / Published online: 11 February 2022

Abstract
The role of a human assistant, such as receptionist, is to provide specific information to the public. Questions asked by the public are often context dependent and related to the environment where the assistant is situated. Should similar behaviour and questions be expected when a social robot offers the same assistant service to visitors? Would it be sufficient for the robot to answer only service-specific questions, or is it necessary to design the robot to answer more general questions? This paper aims to answer these research questions by investigating the question-asking behaviour of the public when interacting with a question-answering social robot. We conducted the study at a university event that was open to the public. Results demonstrate that almost no participants asked context-specific questions to the robot. Rather, unrelated questions were common and included queries about the robot's personal preferences, opinions, thoughts and emotional state. This finding contradicts popular belief and common sense expectations from what is otherwise observed during similar human–human interactions. In addition, we found that incorporating non-context-specific questions in a robot's database increases the success rate of its question-answering system.

Keywords Human–robot interaction · Question-answering robot · Service robot · User study · In-the-wild study · User's experience

1 Introduction

Imagine you at the Louvre Museum in Paris and are being assisted by the museum staff during your visit. Would you ask to the tour guide "Do they believe in God" or "What is the meaning of life"? Unless you wanted to mock the staff at the museum, you would probably ask questions more related to the specific context in which you are situated, say "Where can I find the Mona Lisa painting". However, what if a social robot is designed to assist you during your visit? Would your curiosity lead you to ask these kinds of questions? This study investigated what type of questions people commonly ask a social robot designed to assist them with a question and answering service during their visit at a university's open day[1].

Anticipating human expectations in human–robot interaction is fundamental to the successful design of a social robot, as expectations affect users perception of robot capabilities [1] and attribution of personality to the robot [2]. For example, to design a social robot which can answer any questions from a user, it is important to anticipate the questions a common user would ask. Previously, a large number of studies attempted to understand how humans interact with a service robot in a public space [3–5]. However, only few studies looked into the type of questions people ask to a social robot in a public space [6,7]. Finding an answer to this question can help in improving the success rate of a question-answering (QA) social robot as well as the user experience by providing appropriate and satisfactory answers most of the time.

The success of a question-answering system partially depends on technical aspects, for example, how accurately it

✉ Syed Ali Raza
syedaliraza.ah@gmail.com

[1] Innovation and Enterprise Research Lab, Centre for Artificial Intelligence, University of Technology Sydney, Sydney, Australia

[1] In some other regions of the world, such as USA, this may also be referred to as "preview day" or "open house"

detects speech and how good is the algorithm to match the heard question with the types of questions it is familiar with. One simple approach to match questions is to employ pre-built repositories of question–answer pairs. These repositories can be used to predict the commonly asked questions for the specific environment in which the system is situated, so as to address them in a contextually correct, intuitive, timely and socially natural way [8]. To address more general factual questions from the public, this strategy can be extended to seek for answers through domain-specific ontologies, web crawling or open-domain question-answering systems, such as IBM Watson[2]. However, when a QA system is to be designed for a (physical) social robot, it adds another difficulty layer for questions answering due to the embodiment factor [9–12] and physical presence [13,14], which can affect users' decision-making [15], thus potentially biasing users' question-asking behaviour. Moreover, the novelty effect can play a crucial role [16–19]. Indeed, during an interaction with a humanoid robot, it is possible that out of curiosity a user may not only ask for service-related or context-specific questions, but also ask for questions about robot's preferences, personality and it's behaviours, as has been found with conversational agents [20].

In this paper, we define *"context-specific questions"* as those questions concerning facts, events, people, things and information that are specific to the situated context and role the robot is currently operating for. Additionally, we refer to *"non-context-specific questions"* as all the other questions that do not fit the present definition given for context-specific questions. Furthermore, for the purposes of our study, we define a QA social robot as a robot humanoid in shape and designed specifically to socially interact with humans. A basic dialogue management system to facilitate question answering and the physical presence of this robot is presumed.

Designing an ideal question-answering system which can appropriately address any question from a user is still a present challenge. However, in this paper, it is not our aim to provide a solution to this challenge by proposing a new QA system having better performance. Rather, we believe that by investigating the question-asking behaviour of people interacting with a QA social robot we gather crucial insights to facilitate the design and development of future QA social robots. Therefore, our intention is to collect and analyse question-asking behaviour of users interacting with a QA social robot in a public event, so to gather a better understanding of the types of questions commonly asked by people to a social robot and assist the design of future social robot applications. This is particularly significant to design applications that can maximise users' enjoyment and users' experience. In addition, our intent is to study users' question-asking behaviour in real-world scenarios (as opposed to controlled laboratory observations) and within real-world spaces (as opposed to chatbots and other disembodied conversational agents).

Therefore, our research question is, *"for a QA social robot hosted in a specific environment, what type of questions is a common user likely to ask which should be incorporated in the design of the QA system?"*.

Estimating what information is usually sought by a user is non-trivial. In most cases, it is unfeasible for a system designer to come up with all the sorts of questions which an ordinary user can ask. Therefore, our methodology is based on conducting a user study to collect the questions while a social robot performs the QA service for the real users. However, the novelty of our approach is that a highly accurate QA service is not needed. Instead, a moderately accurate QA service is used to observe users' real-world question-asking behaviour. Moreover, we believe this is the first study where a QA social robot is used in a university's Open Day setting to study users' real-world question-answering behaviour.

In this study, we deployed Pepper robots in a university to answer the queries about the university's open day. The visitors of the open day interacted with the robot at their own will and asked questions of their choice. The users were framed in the situated context of the open day. The robot listened to the questions and tried to provide an appropriate answer by using its QA system. Additionally, we deployed another Pepper robot in a similar setting, except that the users were from one of two user-groups. In one group, the users had participated in an IT tour organised by the university tour guides. At the end of the tour, the human tour guide offered users the opportunity to ask any unanswered questions to the QA social robot. The second group consisted of participants spontaneously interacting with the robot but that had not taken part in the IT tour. All questions asked by the users and the post-interaction survey responses were recorded for analyses.

Through our study, we found that people asked many non-context-specific questions to our QA social robots. Indeed, people asked our social robots many questions about the robots' preferences, beliefs and capabilities. Hence, we found that the success rate of the QA social robot system can be increased significantly by predicting and incorporating non-context-specific question–answer pairs in the robot's knowledge repository. Moreover, participation in a guided tour did not have an effect in prompting the users to ask context-specific questions. Furthermore, the survey results showed users highly enjoyed interacting with the robots. Moreover, the robot's failure to answer a question did not have an effect on users' enjoyment ratings. Note that these findings are limited to situations where a user interacts with a social robot once in a single day public event (like, University Open Day).

[2] https://www.ibm.com/watson, accessed October 2020.

The contributions of this paper are as follow,

- We have presented a method based on deploying a question-answering robot in an in-the-wild setting to collect data on users' real-world question-answering behaviour in a large public event.
- By presenting and discussing our findings (from the analysis of the collected data) we are contributing to assist the design of question-answering social robots for the real-world applications.

In Sect. 2, we explore works related to our present study. Our hypotheses and methodology are explained in Sect. 3. Section 4 discusses the design choices used in our experiments. Section 5 provides the results and Sect. 6 the analyses. We conclude the paper with a general discussion in Sect. 7, and outline our conclusions, limitations and future work in Sect. 8.

2 Related work

Traditionally, many researchers in the domain of natural language processing proposed QA systems based on logical reasoning and linguistic analyses [21]. Other researchers proposed solutions based on semantic web [22,23]. Recently, deep learning has dramatically improved the performance of the QA systems [24,25], especially in the sub-domain of visual QA [26,27]. The voice-assisted devices, like Alexa and Google Assistant, partially solve the problem by answering questions for which the information can be retrieved from the web [28,29]. Nevertheless, any state-of-the-art is far from an ideal QA system.

There has been a substantial amount of work towards making QA systems that deliver specific information requested by a user. These systems are typically built to understand natural language questions and to also reply in natural language. In the field there is relatively more literature available on non-embodied or virtual conversational agents, sometimes referred to as "chatbots", compared to robots. This is mostly due to the technologies earlier maturation and adoption. For example, Kopp et al. [30] provided a detailed study on a virtual-reality-based conversational agent, named "Max", used in a museum in Germany. The human avatar of Max appeared on a static screen and provided visitors with the information about museum. Also, it responded to the user's keyboard-based input with natural, verbal and non-verbal feedback. In a sub-study, they analysed the categories of user-questions and found some frequently asked questions similar to ours. However, unlike our research question, they did not specifically identify the proportions of context-specific and non-context-specific questions. Moreover, the significant difference in our work is that we used an embodied social robot.

More recently, conversational agents have taken the form of personal voice assistants, like Siri, Cortana and Google Now. Luger et al. [31] found via a user study that proprietary conversational agents from many high tech companies failed to meet user expectations due to lack of transparency about the system's capacity and intelligence. Sugiyama et al. [32] developed a conversational agent that incorporated answers to specific personality questions from a dialogue database containing multiple personalities and reported that answering personality-specific questions was effective in preventing dialogue breakdown. This investigation was similar to our study, however, it did not specifically test with a humanoid social robot.

Previous work in HRI for robotic systems involving speech has used human–human interaction as the basis for dialogue. For example, Cantrell et al. [33] used human–human interaction in a team search instructional task to create their dialogue corpus for training. However, for work with humanoid social robots, this may not yield the best dialogue results due to the difference the physical presence of the robot may bring to the conversation [13]. Cruz-Sandoval et al. [34] recognised this issue and proposed the creation of a HRI corpus for conversational dialogue for training machine learning systems. They argue that human–human conversations are insufficient inputs for creation of training dialogue, reasoning that human–human conversation is different to human–robot conversation, due to humans adapting their communication to the social interface and factors such as embodiment and non-verbal language. We concur with this argument. Nevertheless, their work differs from ours due to the nature of the dialogue capture, as they suggest to use a Wizard of Oz (WoZ) system for the human–robot conversation generation. In our system we avoid posing the robot as an intelligent agent, due to the subsequent conversations this may stimulate. Instead, the robot informs users if it cannot answer, and that the answer is not yet in it's database, reinforcing it's limited robotic nature and limiting subsequent adaptions users may make if it was considered of greater intelligence.

Similarly related work to ours which uses social robots in the public to capture natural, human–robot interaction dialogue, without use of a WoZ, is sparse. Ben-Youseff et al. [35] have made publicly available a data set of spontaneous interactions with a Pepper robot that contains dialogue. This however was created for the purpose of studying and predicting engagement by users and does not provide a question–answer scenario, as used in our study.

Lee et al. [6,7] used a receptionist robot, roboceptionist, to analyse human–robot dialogues. Their initial study [6] showed that human–robot interaction varies from human–human interaction, and giving the robot a personality with a background and occupation can provide a common ground for conversation. In a subsequent study [7], the authors showed that the user's choice to greet or not to greet at the start of an interaction can predict if the user will perceive the robot as an information point or a human-like receptionist. They found that the users who greeted the robot perceived it more as a person (i.e. more robot-related and person-related topics were discussed) and those who did not greet treated the robot as an information kiosk. Our study differs from the roboceptionist's studies in the embodiment of the robot, QA system design (single question answering vs. dialogue scripts), duration of data collection (single day vs. weeks of data) and the context (university's open day vs. no specific occasion). The roboceptionist had no arms, a monitor with an avatar of human-like face, computerised voice and keyboard-based question input. Also, the robot was confined behind a desk. On the contrary, Pepper is a state-of-the-art social humanoid robot designed to interact with people in a social and intuitive way, by listening and replying with voice, gestures, head movement and eye contact.

Another similar study was done by Bohus et al. [36] at the Microsoft Research Center at Richmond using a direction providing NAO robot which spanned over a week. Similar to our findings they found that only a small percentage of interactions were need-based engagements (i.e. genuine direction-seeking interactions) and a majority of interactions were driven by curiosity to test the robot's ability. Our study differs in that we have used a more advanced social robot, our context is a one-day event and our users interacted with the robot only once. These "first-time" users are those typically expected from social robots deployed to provide services in a public space. As such, our collected data can better assist the design of such social robot applications in public spaces.

Guo et al. [37] have performed a study most similar to ours. They built a learning QA system on a Pepper robot using IBM Watson Natural Language Classifier, trained to the specific context of a robot concierge within their institution's "ThinkLab". While their technical system has parallels to ours, their study focuses on the ability of their system to learn new classes of questions, not on the type of questions to be expected from users "in the wild". Indeed, our study differs in that we specifically had participants who were freely motivated to spontaneously interact with our QA system, with no constraints on what questions they could ask.

Indeed, our impetus for collecting questions has been not only to create a knowledge repository of non-context-specific questions, but also to craft appropriate answers that assist creation of an enjoyable experience with a QA social robot. Previous work suggests for HRI design and social acceptance that creation of enjoyment expressly be considered, not just ease of functional use [38].

3 Hypotheses and method

3.1 Hypotheses

Our objective is to investigate our research question by evaluating three hypotheses.

Firstly, from our previous in-the-wild studies investigating social robots applications deployed in public spaces we observed that people asked a non-negligible amount of non-context-specific questions to the robot. For example, when we deployed a Pepper robot, enabled by a Wizard of Oz (WoZ) style question-answering service, at the entrance of our university's main building or when we deployed a Pepper robot, enabled by a question-answering service very similar to the one used in this study, to answer questions by the visitors of RoboCup 2018[3], we found a majority of questions were non-context-specific. Therefore, we base our first hypothesis on this observation.

H1:A significant amount of questions asked to a QA social robot, deployed in a public space, would be non-context-specific.

Secondly, previous works show that people's interactions with robots can be driven by the curiosity to test the robot's abilities [36] and that when the robot is perceived as a human-like agent, the dialogues between people and such robot include topics concerning the robot's preferences, events, beliefs and thoughts [7]. Therefore, if the amount of non-context-specific questions asked by people to a QA social robot is indeed significant, we will expect to see that a QA social robot designed to answer such questions would have a significantly higher success rate as compared to a QA social robot only targeting context-specific or service-related questions.

H2: For a QA social robot, being able to answer non-context-specific questions will significantly increase the QA system's success rate (i.e. the ratio of correct to incorrect answers).

Thirdly, we believe that in general it is fair to expect that people who encounter the robot within a frame of participation in a specific activity will ask more questions concerning such activity or context. For example, people taking part in a tour of a museum would more likely ask the (human) guide questions that remained unanswered during the tour or additional information concerning the museum; especially if the guide invites them to. Should we expect a similar behaviour if the framed users are offered to ask questions to a social robot instead?

H3: Users interacting with a QA social robot after participating in a specific framing activity are more likely to ask context-specific questions, as compared to users not attending such activity.

[3] http://2018.robocup.org/, accessed October 2020.

3.2 Method

To test the above hypotheses, we employed a simple methodology. Our method is based on designing a functional question-answering service for a social robot and deploying it in a real-world scenario to collect users' questions and performing post-study analysis on the collected data to determine the types of users' questions. Therefore, we first designed a question-answering service for the Pepper robot. We ensured the robot was able to a) perform the start-to-end interaction with the humans and b) answer some of the users' questions. Note, designing a highly accurate service is not needed in our methodology. The implementation details of the service are provided in Sect. 4.2.

Next, we deployed the robot to use the service with the end-users under the real-world settings. In our methodology, the control over a user's interaction with the robot was kept to a minimum. This is in contrast to the typical laboratory settings, which are usually controlled and based on strict protocols for the users. As mentioned in Sect. 4, the robots were hosted in publicly accessible places and a simple instruction was given to the users before the interaction.

Finally, the users' questions were analysed in a post-study exercise.

4 Experimental design

We conducted our experiment during the University of Technology Sydney (UTS) open day. Specifically, our robots were exhibited in the Faculty of Engineering and Information Technology (FEIT) building.

During the open day at FEIT building, the university offered several information sessions for FEIT degrees. In addition, research groups exhibited technologies available at the FEIT building and offered workshops on IT and Engineering topics. The faculty also offered several guided tours during the day: IT tours and Engineering tours.

The study and the methodology discussed in this paper were approved by the UTS ethics committee.

4.1 Robot platform

For our study, we employed three Pepper[4] robots (Fig. 1). We placed the first Pepper robot (Robot$_{greet}$) at the entrance of FEIT building. This robot welcomed people by waving at them and promoting the university and its open day activities. The second robot (Robot$_{exhibit}$) was placed on the exhibition level, together with other technologies and interactive experiences available for the audience. Finally, we placed the third robot (Robot$_{tour}$) in a classroom situated on a different level. This classroom was also the last stop of the IT tour.

We did not employ Robot$_{greet}$ to collect data for our experiment. Indeed, this robot focused on promoting Robot$_{exhibit}$ on the exhibition level and the IT tour (people ended their tour by interacting with Robot$_{tour}$). Therefore, Robot$_{greet}$'s goal was to maximise the number of interactions between the guests and the other two robots, so to increase the sample size of the collected questions.

Robot$_{exhibit}$ area was surrounded by colourful display banners inviting people to interact with the robot (i.e. "Meet a robot" and a large cardboard banner of a standing Pepper robot). People had to queue in order to access the robot for a one-to-one interaction.

We placed Robot$_{tour}$ in a separate classroom at the end of the IT tour to perform an additional analysis, investigating if a different framing context can impact on the type of questions asked by people to the robot. People were able to either pop in the classroom at any time and wait for their turn to interact with the robot, similarly to Robot$_{exhibit}$, or they visited the classroom after taking part in the IT tour. When each IT tour was about to conclude, the classroom was made inaccessible to other members of the public so as to prevent interactions between the two groups.

Robot$_{exhibit}$ and Robot$_{tour}$ offered the very same question-answering service and used the same tablet user interface to collect participants' survey responses. An example of human–robot interaction setup used in this study is shown in Fig. 1. The same interaction setup was used for both Robot$_{exhibit}$ and Robot$_{tour}$.

4.2 Question-answering service

The question-answering service provided by Robot$_{exhibit}$ and Robot$_{tour}$ consisted of the following steps: (i) listening to the question of the participant; (ii) processing the question with Google Speech-to-Text cloud service[5] to obtain the text transcription of the spoken question; iii) looking up for the question on the robot's knowledge base so to rank the entries by similarity; iv. A) directly answering to the participant by using the answer from a question in the knowledge base having highest similarity, if above a certain threshold, or iv. B) telling to the participant that a perfect match for that question was not available in the knowledge base, but that an answer to a closer question was found, if below the threshold[6]; v) asking to the participant if he/she is willing to ask another question. If the participant answered "Yes" the process repeated from step (i), otherwise the interaction ended by thanking the participant and asking them to answer a quick survey question.

[4] https://www.softbankrobotics.com/emea/en/pepper, accessed October 2020.

[5] https://cloud.google.com/speech-to-text/, accessed October 2020.

[6] In this situation, the robot waited for the participant's confirmation before proceeding answering

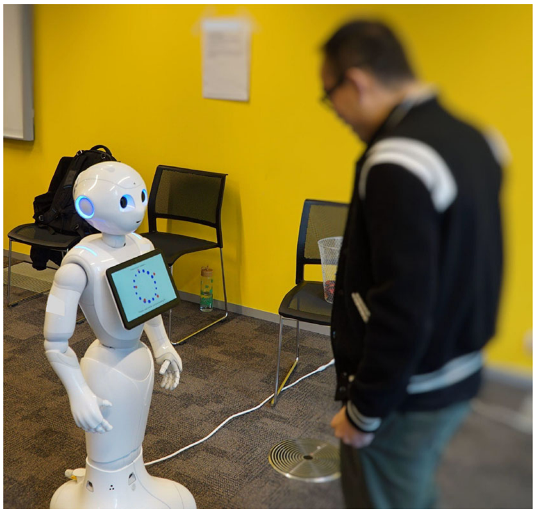

Fig. 1 An example of how users interacted with the robot. The users stood in front of the robot at a convenient distance (which they decided by themselves). The tablet attached to the robot is used to display animations to represent robot's internal state (such as "listening", "processing" or "speaking"), alongside captions to help user read what robot just said (in case the user did not hear properly), and button(s) to record input about a question.

Note, the robots used synthetic voice to produce speech in steps iv and v.

To collect the participants' questions, we employed audio recorders stuck on the back of the robot's tablet. People were informed about their audio being recorded and they understood that by interacting with the robot they provided consent for collecting their audio recordings for our study. However, participants did not know the final aim of the study. At the beginning of every interaction, the robot welcomed the user and repeated the current time in hours, minutes and seconds. We used this information to assign a unique ID to the current session/participant (e.g. the speech "12 hours, 35 minutes and 36 seconds" became the session ID 123536). We mainly used this information while transcribing the audio recordings from the external microphone. It facilitated us in determining which question in the recording belongs to which user. We also assigned a progressive unique question ID to all the questions correctly recorded during the interactions.

To adhere to our development approach of short, test and learn iterations, and with the goal of software agility, we constructed a knowledge repository of common questions and their answers, and employed a simple matching mechanism. Over a series of previous robot demonstrations in several environments and occasions, we have noted common questions asked by members of the public. We then manually compose answers to those questions and identify key-words that might also be appropriate to trigger that response. When the robot is asked a question, the question is transcribed using Google Speech-to-Text and that output is converted into a phonetic form using the Double Metaphone algorithm [39]. A minimum edit distance metric is used to find a high-similarity question in the knowledge repository and the corresponding answer is used. Where no similar question can be found, the system defaults to a simple match based on the presence of identified key-words. It is important to remark that this study is not aiming to measure the performance of our present question-answering service by comparing it with other available algorithms. Instead, the objective of this study is to acquire information on the type of questions commonly asked to a QA social robot by the public, so to suggest directions for boosting its success rate and increasing user's experience, enjoyment and satisfaction.

4.3 Experimental settings

To investigate the type of questions that our participants asked to the robots, we did not employ any independent variables. The question-answering application was the same for both Robot$_{exhibit}$ and Robot$_{tour}$ and the researchers introduced both the robots as an opportunity to ask them any question they wanted to ask and getting an answer to their questions. The instruction given to the users of Robot$_{exhibit}$ was "ask any questions to the robot". Each participant interacted once with either Robot$_{exhibit}$ or Robot$_{tour}$.

People attending the IT tour and interacting with Robot$_{tour}$ were offered the chance to ask any question they wanted to the robot as well. However, the human tour guide instructed the users at the final stop of the tour (i.e. the classroom with the robot) "ask to the robot any questions that you have had, instead of asking such questions to me (the human tour guide)". "*Tour*" vs. "*no tour*" was the independent variable we used for testing our third hypothesis (H3).

4.4 Participants

The participants of our study were guests visiting the FEIT building during the university's open day. They were free to interact with the robot and signage surrounding the robots warned the participants that by interacting with the robot their audio would be recorded and that by interacting they give the consent to collect non-identifiable audio for a research study, as per the methodology approved by the university's ethics committee. Additionally, participant information sheets were provided.

The demographic at the event was composed in the majority of young adults and their parents. For this study, we did not collect the age groups of our participants, but only their gender, which was annotated by the researchers while listening and transcribing the audio recordings.

A total of 85 participants took part in our study. For some of them, the audio devices failed in recording their questions, but the QA system was able to correctly transcribe the

Table 1 Gender distribution of the participants of our study.

Gender	Robot$_{exhibit}$	Robot$_{tour}$		**Overall**
		tour	**no tour**	
Male	32	10	7	49
Female	24	3	4	31
N/A	5	0	0	5
Total	61	13	11	85

questions by mean of Google Speech-to-Text service. We annotated the gender of such participants as "not available" (N/A). Table 1 summarises the allocation of our participants by robot and, for Robot$_{tour}$, the allocation in either "tour" or "no tour" conditions. Note that we have provided gender distribution in Table 1 to show that the sample was not biased towards any gender. For example, there were 39% female users in case of Robot$_{exhibit}$ and 29% female users in case of Robot$_{tour}$. It shows that the representation of both genders was though not balanced but also not biased.

4.5 Measures

In this study, we measured: (a) the type of questions asked to our robots and (b) the participant's enjoyment for using the provided service.

4.5.1 Question type

To assign the questions to either the class "context-specific" or "non-context-specific", we first assigned each question to a specific question template. One of the researchers of this study, generated 98 distinct question templates by observing the questions asked by the users. The templates were designed to fit the various forms of the same query. The criteria for grouping queries within a single template were the following:

1 The queries assigned to a single template can all be acceptably answered with the same answer;

2 A query is assigned to a pre-existing template only if the query can be assigned to it after adapting the pre-existing template with optional insertions or exclusive alternatives that do not dramatically alter the structure of the pre-existing template.

To better understand rule 2, consider the following examples. The query "Hello, how are you?" can be initially assigned to the template "hello how are you". If a new query "how are you?" is observed, this new query can fit the previous template by adapting it with an optional insertion, thus becoming "[hello] how are you", where the square brackets indicate an optional part of the template. Similarly, the query "Where is classroom?" can be initially assigned to the template "where is classroom". When the query "How can I get to classroom?" is observed, the pre-existing template can be adapted with an exclusive alternative, thus becoming "(where is | how can I get to) classroom", where the pipe character separates exclusive valid alternatives for the considered template.

Also, note that the only exception to rule 1 is if the queries assigned to the same template will require different answers but these answers can be exhaustively computed by a single function. This is the case of simple mathematical questions like "What is 1 + 1?" or "what is the square root of 2?" fitting the template "what is [the] $math_formula". This template can be answered by parsing the mathematical formula from the query and returning its solution.

After having each question assigned to a template, each template was assigned to a code describing the nature of the query. This coding technique is similar to the methodology proposed by Mutlu and Forlizzi [40] and is based on the grounded theory [41–43]. The researcher who generated the templates, also identified four codes that were able to classify the templates in mutually exclusive groups. Specifically, these codes helped in identifying if the question asked was to seek context-specific information or not. Below are the descriptions for these codes:

1 Dialogue openings / closings (DOC): Expressions that are used to open and close a dialogue and they do not really mean to ask for an information or *they do not expect an answer in return* (i.e. commonly accepted expressions to start or end the conversation).

- Example: *[hi] how are you [going] [today]*

2 Role oriented (RO): Queries seeking information or opinions expected by the professional role of the robot in the considered context. That is, for the university study, queries asking information about the university, the life at the university, *opinions about the university and other universities*, the spaces at the university and information to move within the university during the event. For the hospital, queries asking for directions, for services inside and in the immediate proximity of the hospital, seeking assistance to find or contact patients or seeking medical assistance.

- Example: *how old is this building*

3 Factual knowledge (FK): Queries asking for information that can be retrieved online (for example from Google), through computations or that is of common knowledge and not directly expected by the professional role of the robot in the considered context.

- Example: *what is the closest planet to earth*

4 Robot directed (RD): Queries asking for information or opinions not expected by the professional role of the robot and that can only be answered by consulting a local knowledge base of the robot containing information about the robot, everyday robot's activities, its opinions, its beliefs, *its skills, its features, its social connections, its preferences, its motivations, its goals, its mental states and its drives*. This code is also used to classify commands, requests or permissions made by the user to the robot that are not expected by the professional role of the robot in the considered context.

  - Example: *where were you (made | born)*

Other three researchers of this study were asked to independently code the templates with the four identified codes. We then tested the coding consistency by performing a Cohen's κ analysis on the obtained codings. There was a substantial agreement between the three researchers, $\kappa = 0.725$, $p < 0.001$. In fact, only 26 templates of the generated 98 were coded differently by the research team. We resolved the disagreements for the 26 templates with a discussion and, as a result, we adapted the criteria to resolve the final misinterpretations and reach an agreement. The adaptations made to the criteria are stressed in italics in the list presented above.

These four codes helped in identifying the types of questions asked to our robots. Only the questions coded as RO were considered as context-specific. Rest of the three codes belonged to non-context-specific class of questions. Within the non-context-specific class, we have further division. The questions coded as FK are named "Googleable" and the questions coded as DOC or RD are named "Others".

4.5.2 Participant's enjoyment

We measured the participant's enjoyment by mean of a single question. We asked: "Did you enjoy interacting with me?" We collected responses using a five-item Likert scale: (1) not at all, (2) very little, (3) neutral, (4) somewhat, (5) to a great extent.

We decided to employ this single question instead of more accurate but longer questionnaires measuring dimensions of users' satisfaction for at least two reasons. Firstly, as already mentioned before, the main objective of this study is to collect, classify and analyse questions that people ask to QA social robots. Thus, we wanted to gauge only a preliminary understanding of people's overall enjoyment to understand if our application was mainly leading to a positive or negative user experience. Secondly, the nature of the environment in which the study was situated discouraged us to involve our participants in thorough questionnaires. Our participants were guests of the open day and it was not our intention to prevent them from spending their time in other activities by instead asking to fill time-consuming questionnaires.

Table 2 Questions asked by our participants

Type of Question	Recorded	Attempted	Correct
Context-specific	15 (8.38%)	12	3
Googleable	43 (24.02%)	37	14
Others	121 (67.59%)	103	65
Total	179 (100%)	152	82

4.6 Inclusion criteria of collected data

4.6.1 Question type

To measure the type of questions asked by people to our social robot, we included in the analyses any question asked by the participants to the robot that was either recorded with our audio devices or correctly transcribed by Google Text-to-Speech. That includes also: (a) those questions that were asked to the robot while the robot was not ready to listen yet; (b) those questions that the robot was not able to process with Google Speech-to-Text (i.e. Google Speech-to-Text did not return a transcription within a timeout) or (c) those questions that the robot was not able to process because the application, or the robot, crashed while attempting in processing them. However, questions falling in category (a), (b) or (c) were excluded when measuring the accuracy of our question-answering application, since those questions were not processed by the system and we do not have data to decide if the robot would have been able or unable to correctly answer them with our knowledge repository.

4.6.2 Availability of participant's feedback

We included in the analyses of participant's enjoyment only those participants that answered the post-interaction question necessary to measure their enjoyment, namely 77 participants out of the 85 that took part in the study.

5 Results

From our 85 participants, we recorded a total of 179 questions. Our robots were able to process and attempt answering 152 questions and 82 of them (53.94%) received a correct answer. Table 2 summarises the type of questions collected from our participants and the system's success rate.

Note, an answer was considered correct if a) transcription by Google Speech-to-Text was considered correct (however, not needed to be exactly the same) b) the transcribed question

Table 3 Examples of correct and wrong attempts.

From Audio Recording	Google's Transcription	Best Match from Database	Correct
How old is this building	How are you feeling	How are you feeling	No
who invented you	who invented you	Who owns you	No
[robot_name] what's the meaning of life I mean real life	[robot_name] what's the meaning of life I mean real life	What's the meaning of life	Yes
Where are we right now	Where are we	Where are we	Yes

and the best matched question from the database were similar. The researchers of this study performed these checks by going through the logs in an ad hoc manner. A simple rule was followed while making these checks. If the context of the question from audio recording and the question from database is the same then it was marked correct, otherwise incorrect. Since the answers to each question in the database were added by the researchers, it was implied that if the robot understands a question correctly then the answer it provides should also be correct, and vice versa. Some examples of marking an attempt as correct or incorrect are provided in Table 3. For example, in the second example in Table 3, the question from the audio recording, "who invented you", had a different context from the question fetched from the database, "who owns you". On the other hand, in the third and the fourth example in Table 3, the best matched questions from the database had some words missing compared to the audio recorded questions, nevertheless the missing words did not change the context of the questions. Therefore, those matches were considered correct.

We further grouped the questions by their templates. This allowed us to investigate the most commonly used templates during our study. Table 4 lists the identified templates, grouping under the template "singletons" the questions having a topic asked only once during the study.

Importantly, only two of the questions classified as context-specific by the assessors were asked twice. Similar questions are marked by a number in superscript in the list below. Rest of the context-specific questions were asked a single time during the study (i.e. they were all singletons). Those questions were:

1 Does UTS have scholarships?
2 Where is the toilet?^{1}
3 How old is this building?
4 where is the bathroom?^{1}
5 is UTS a good university?*
6 where is the closest staircase?^{2}
7 where's the closest staircase?^{2}
8 what is engineering?
9 what's the answer to HEC questions?
10 is [name of another Australian university] a bad university?
11 where is UTS?
12 what do you know about UTS?
13 Where are we right now?
14 How old is UTS?
15 What do you like about UTS?*

We can notice that even within these questions annotated as context-specific, we still have some questions asking to the robot to express its beliefs or preferences. Such questions are marked in the above list with an asterisk (*).

Furthermore, in Table 5, we provide a summary of the collected questions excluding dialogue initiators or terminators (DOC). We will use these additional results for further analyses in Sect. 6.1.

We counted the number of questions that each participant asked to the robot during a single interaction. The majority of people (59, 69.41%) asked only one or two questions before ending the interaction with the robot (respectively, 38, 45%, and 21, 25%, participants). As an upper bound, 7 was the maximum number of questions asked by a single participant. This frequency level was reached only once during the entire study. Figure 2 depicts the distribution of number of questions asked by each single participant.

Table 6 summarises the frequencies of each type of question limited to the interactions with Robot$_{tour}$ and divided by tour condition (i.e. tour *vs* no tour). Table 7 presents the number of participants interacting with Robot$_{tour}$ and asking at least a context-specific question divided by tour condition. Even in this case, the proportion of questions for each type seems to be aligned with what we found for the overall study, namely that only a small percentage of people asked context-specific questions to our QA social robot. In "tour" condition two participants and in "no tour" condition only a single participant asked a context-specific question to the robot.

Finally, Table 8 shows the descriptive statistics for the perceived level of users' enjoyment, as per scores collected from 77 of the 85 participants taking part in this study. The table

Table 4 Questions asked, ranked by templates

Rank	Question template	#	%
1	[hi | hello | so] what (is | s) your name	22	12.29
2	[hi] how are you [going] [today]	9	5.02
	what (is | s) $math_formula	9	5.02
3	how old are you [banana]	8	4.46
4	(what | how) (is | s) the weather [of today | today]	7	3.91
5	can (we (do handshake | shake hands) | you shake (my hand | hands [with me]) | I shake your hands | shake my hand)	6	3.35
6	who ((invented | made) you | is your creator)	5	2.79
7	(what | which) is your favourite colour	4	2.23
	[$robot_name] (what is the | is there a) meaning of life [I mean real life]	4	2.23
8	what (is today's day | day is today)	3	1.67
	who is (the prime minister of australia | australia's prime minister)	3	1.67
9	tell me a joke	2	1.11
	([can you please tell me] where is the | I want to go to the) [public] (toilet | bathroom) [in the ground floor] [please]	2	1.11
	how many languages [do] you speak	2	1.11
	where (is | s) the closest staircase	2	1.11
	how tall are you	2	1.11
	where were you (made | born)	2	1.11
	(can I have | give me) a high five	2	1.11
	do you like your (team | creators)	2	1.11
	who are you	2	1.11
	[can you] dance	2	1.11
	do you believe in god	2	1.11
	(do you think its is going to | will it) rain today	2	1.11
11	Singletons	75	41.89

Table 5 Questions asked by our participants (excluding dialogue openings and closings)

Type of Question	Recorded	Attempted	Correct
Context-specific	15 (8.82%)	12	3
Googleable	43 (25.29%)	37	14
Others	112 (65.88%)	94	57
Total	170 (100%)	143	74

offers the statistics separated in two groups: (i) participants that received at least a wrong answer by the robot and (ii) participants that did not receive any wrong answer by the robot during their interaction.

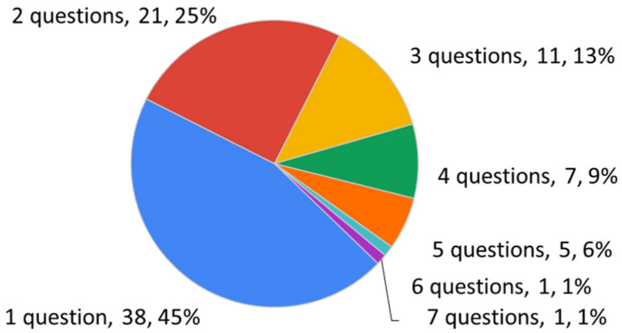

Fig. 2 Number of participants grouped by the number of asked questions during their interactions

6 Analyses

In this section, we will provide the analyses to our results by looking at the type of questions collected, the success rate of the discussed QA system and the level of enjoyment estimated from the post-interaction question.

6.1 Question type and success rate

From Table 2, we can clearly see that people did not ask many context-specific questions. Indeed, we collected only

Table 6 Questions asked by our participants when interacting with Robot$_{tour}$

Type of Question	Tour	No tour
Context-specific	2 (8.33%)	1 (4.35%)
Googleable	7 (29.17%)	6 (26.08)
Others	15 (62.5%)	16 (69.56%)
Total	24 (100%)	23 (100%)

Table 7 Number of participants interacting with Robot$_{tour}$ and asking context-specific questions in the framing study conditions

Did the participant ask a context-specific question?	Tour	No tour
Yes	2 (15.38%)	1 (9.09%)
No	11 (84.61%)	10 (90.91)
Total	13 (100%)	11 (100%)

15 context-specific questions over a total of 179 questions, which equals to only a 8.38% of the total collected questions. There was a significant difference between the proportions of context-specific and non-context-specific questions asked to the robot ($\chi^2 = 124.03$, $p < .00001$)[7]. Even by excluding the dialogue initiators and terminators from the study, as summarised in Table 5, we only reach a 8.82% frequency level for context-specific questions. The difference still remains significant between the two proportions ($\chi^2 = 115.29$, $p < .00001$)[8]. Therefore, our first hypothesis (H1) that a significant amount of questions (approximately 91% of the questions) asked to the robot are non-context-specific is accepted.

Let us assume to have an ideal QA social robot able to answer any context-specific question that was correctly processed and attempted by the robot with a 100% success rate. Let us also assume that this ideal system can also answer any other non-context-specific factual question (i.e. FK) (e.g. "What's the colour of the sea?"). Assume also that this ideal QA social robot was not designed to answer other questions like robot-directed ones (i.e. RD), such as personal preferences and opinions (e.g. "What is your favourite colour?"). We denote this ideal QA system with "QA service oracle". This ideal QA system is depicted as a service robot in Fig. 3. In addition, let us assume that the population sampled in this study is a representative subset of the whole population of users interacting with a QA social robot (at least at the present moment in time). Given those assumptions, we can measure the performance of such ideal QA service oracle by counting all the context-specific and Googleable questions collected in this study and attempted by the robot in the present study as *hits*, and all the other remaining questions asked and attempted by the robot as *misses*. By doing so, in our scenario this ideal QA system can only achieve a 32.23% success rate (12 + 37 = 49 over 152 questions). Now, let us consider our QA social robot designed in such a way to include in its knowledge repository also robot-directed questions and other non-context-specific questions that we included from observed human–robot interactions in previous studies. Such a QA robot is depicted as a social robot in Fig. 3. By considering the questions for which the QA system attempted an answer in the present study, we achieved a success rate of 53.94% (82 over 152 questions), which is 1.67 times the expected success rate for the ideal QA service oracle. Even when excluding the dialogue initiators from our study (see Table 5), the ideal QA service oracle can only achieve a 34.26% success rate (49 over 143 questions), whereas our designed QA social robot achieved 51.74% success rate (74 over 143 questions). We conducted a pairwise proportions z-test analysis testing the success rate of the ideal QA service oracle against our actual QA system's success rate. We found that the two success rates significantly differ ($p < 0.001$ including DOC and $p = .00278$ when excluding DOC), thus validating our second hypothesis (H2).

As discussed in Sect. 5 that 38 users asked only one question during their interaction with the robot. One may wonder why would a participant new to the robot would ask only one question? Therefore, we analysed those single questions. We found that half of them (19 out of 38) were answered correctly. Therefore, this user behaviour cannot be attributed to the accuracy of the QA service because 19 users ended the interaction even after having the correct answers for their questions. Moreover, 11 of them were dialogue initiators, and none of them were contextual questions. It shows that those participants were most likely not interested in asking contextual questions. We believe those participants wanted to test the robot's technology for which asking a single question was sufficient. Note that the process of asking questions and getting answers was identical for any number of questions.

6.2 Tour guide framing effect

Given our small sample for the framing study, we conducted a Fisher's Exact test to compare the two considered conditions (tour vs no tour) as per results presented in Table 7. The relation between number of participants asking at least one context-specific question and tour framing condition was not significant, $p = 1.000$. Therefore, we do not have enough evidence to validate our third hypothesis (H3), namely that framing participants with a specific activity within the context in which the robot is employed can lead to more context-specific questions from users.

[7] Chi-square goodness of fit test was used. Observed frequencies were 15 and 164 and the expected frequencies were both 89.5.

[8] Observed frequencies were 15 and 155 and the expected frequencies were both 85.

Table 8 Descriptive statistics of users' enjoyment

	(i) With misses	(ii) Without misses	(i + ii) Overall
Number	52 (67.53%)	25 (32.47%)	77 (100%)
Mean	4.154	4.44	4.247
Median	4	5	4
StDev	0.802	0.87	0.83

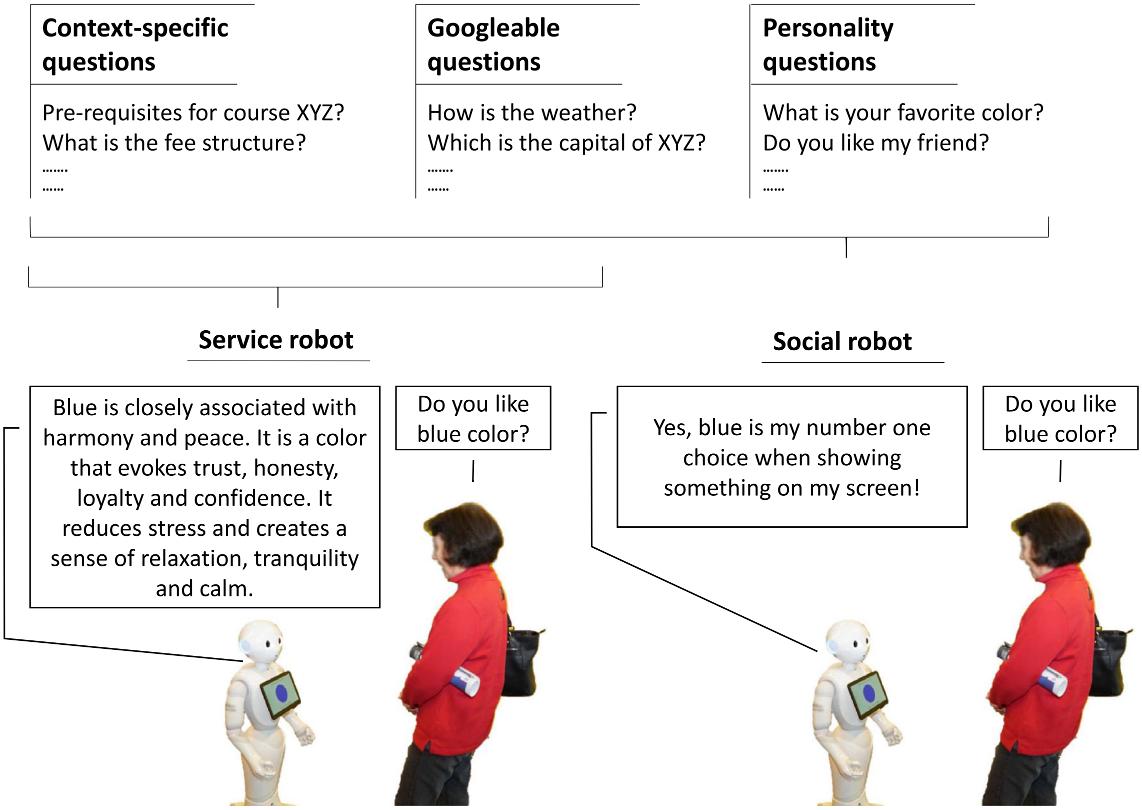
Context-specific questions
Pre-requisites for course XYZ?
What is the fee structure?
....
......
Googleable questions
How is the weather?
Which is the capital of XYZ?
....
......
Personality questions
What is your favorite color?
Do you like my friend?
....
......
Service robot
Social robot
Blue is closely associated with harmony and peace. It is a color that evokes trust, honesty, loyalty and confidence. It reduces stress and creates a sense of relaxation, tranquility and calm.
Do you like blue color?
Yes, blue is my number one choice when showing something on my screen!
Do you like blue color?

Fig. 3 At the top, showing the three categories of questions considered in our analyses. At the bottom, showing two scenarios, a service robot which replies by finding the fact from the web, and a social robot which replies as per the design of its personality. The service robot's reply shown is the top answer from Google's featured snippet API for the query "do you like blue color?" (accessed on October 2020).

6.3 Participant's enjoyment

By looking at Table 8, we see a positive trend on participant's enjoyment level. From a scale of 1 (not at all enjoyed) to 5 (enjoyed to a great extent) the means of all the considered groups exceeds the score of 4. We compared the level of enjoyment of those participants that experienced at least a system's miss during the interaction with the robot (column (i) in Table 8) and those that did not experience any system's miss (column (ii) in Table 8) by employing a Matt–Whitney test for unpaired data obtaining a p value of 0.0634. Therefore, we do not have enough evidence to demonstrate that failure in answering users' questions significantly decreases the perceived level of users' enjoyment of the robot application. Nevertheless, the p value indicates the presence of the "marginal effect"[9].

7 Discussion

Our results presented in Sect. 6.1 validate our first hypothesis (H1). As expected, a very large number of questions asked to the robot were non-context-specific. The results presented in Sect. 6.1 also validate our second hypothesis (H2), namely that a QA social robot predicting and correctly answering also non-context-specific questions (which

[9] As pointed out in Reference [44], the "marginal effects" are also important in human–robot studies. They are referred to as "trends", i.e. when $0.05 < p < 0.1$.

includes robot-related questions) would experience a significant boost on the success rate of its QA system. In fact, the majority of questions asked by the university's visitors to the QA social robot were requests for the robot's opinions and preferences. Only a very small number of them were context-specific or answerable by an intelligent system like Google Web Search or IBM Watson. In addition, some of the questions assessed as context-specific by the independent assessors asked the robot for opinions or judgements. By comparing the expected success rate of an ideal QA service oracle to our actual implementation we found that including non-context-specific questions in the question–answer pairs repository of a QA social robot can significantly increase the success rate of the QA system.

We believe these results will remain valid for environments similar to a university's open day. These are the environments where people come to explore or entertain. For example, if a QA social robot is deployed at the reception of a conference or at a product's promotional event, or at the entrance of a concert, one can expect similar question-asking behaviour from the users (i.e. asking mostly non-context-specific questions). On the other hand, if the robot is situated in an environment where the need for a service is urgent or highly desirable, then this type of question-asking behaviour by the users will not be expected. Instead, in those scenarios, we would expect the majority of questions to be context-specific. For example, if a QA social robot is deployed at the reception of a hospital, or at the counter of a visa centre, then it is likely that users would ask mostly service-specific questions.

It is important to note that the results reported in Sect. 6.1 differ from the results reported in some of the previous studies [6,7]. Lee et al. reported higher rates of context-specific or information seeking questions (40-50%), however, with no mention whether those were asked in the first or one of many interactions. It is likely that the users had overcome the curiosity about the robot, if they had any, in the initial interaction(s), and subsequently they interacted only when they needed some information. However, the details of the first and the subsequent interactions were not discussed separately in the work of Lee et al. [6,7]. Moreover, the higher rate of context-specific questions can be attributed to their coding of information seeking or instrumental questions which included the information about weather, date and time, in addition to the information about location, event, person, which is not the case in our study. Nevertheless, a majority of questions in those studies were non-context-specific, which highlights the need to answer such questions to increase the success rate of a QA social robot. Finally, QA social robots deployed in public spaces are expected to encounter many "first-time" users visiting the space and the robot for the first time and, as such, they must be able to predict non-context-specific questions asked by users out of their curiosity.

Additionally, our results from the analysis in Sect. 6.2 suggest that even framing users by situating them in a more context-specific activity, it seems to not impact on the questions people ask a QA social robot. However, a further investigation in this direction is needed in order to test if that is the case for different framing contexts. Nevertheless, we believe situating the QA social robot in a more service-focused environment will provide better framing to prompt the users to ask more context-specific questions.

If we look at the achieved success rate within each single class of questions, although our system was able to achieve a 63.10% success rate (65 over 103) for the non-context-specific questions, we achieved only the 37.83% of success rate (14 over 37) for Googleable questions and a dramatic 25% success rate for context-specific questions. Failure to answer most of the context-specific questions is obviously one limitation of this work. However, the main reason for this limitation is the difficulty in anticipating such questions.

When building the context-specific knowledge repository, we involved organisers of the university open day as knowledge domain experts. Hence, it should be expected that we would have been able to answer at least some context-specific questions. Instead, these results convey, that even by using domain experts for predicting the context-specific questions commonly asked by people in a specific domain, it is quite difficult to include the questions that the robot would actually encounter when performing it's QA service. This difficulty derives from the fact that such context-specific questions are: (i) *very limited in number* during each testing iteration as compared to non-context-specific questions, thus extending the knowledge base only to a limited extent when using an Agile and User-centred design approach [45], (ii) *very diversified topics*, with each topic occurring only once during a full test iteration (see the list of the collected context-specific questions in Sect. 5), thus being highly unpredictable even by domain experts, (iii) *composed of hard to predict grammar*, conversely to other, robot-directed questions, which follow a more structured grammar derived by social dialogue norms and more easily predictable (e.g. "How are you?/How are you doing?" or "What's your favourite colour?/Which is your favourite colour?"). Additionally, it indicates that users' context-specific questions to a QA social robot can be different from the context-specific questions users ask to a human service-person. Therefore, knowledge domain experts might not be able to tell what type of questions users would actually ask to a QA social robot. However, more experiments are needed to understand the difference between users' context-specific questions for a QA social robot and a human service-person.

In our experiments, people were able to ask any kind of question independently of knowing if the robot was actually capable or not to answer such question, and they were not aware of our research methodology and questions categorisation (context-specific vs non-context-specific). Therefore, it is unlikely that their question-asking behaviour switched from asking context-specific to non-context-specific questions during the interaction for reasons like robot failures or the provision of unsatisfying answers.

The available literature [32,36] and our previous experience in deploying social robots applications for in-the-wild studies, conferences and other events, has suggested that people ask a non-negligible number of non-context-specific and robot-directed questions. Hence, our question–answer pairs repository had a large number of these type of questions and ad hoc answers. This approach led our system to achieve about 54% accuracy in an in-the-wild study during a public university's event.

Finally, as expected, when looking at participant's enjoyment we found that on average people provided a higher score if they did not experience any system's miss during the interaction as compared to participants that experienced at least one system miss. However, the provided analysis did not find enough evidence to demonstrate that such difference was significant. Additionally, it must be noted that asking the user to answer the survey on the robot may have increased positive responses, as per Media Equation theory [46] and more recently Hoffman et al. [47]. Regardless, the overall enjoyment was rated positively, with a more positive trend for users that did not experience robot's failures.

However, we believe that by increasing the sample size and by employing more comprehensive questionnaires investigating several dimensions of users' enjoyment and satisfaction we might be able to observe an effect between QA system's failure and dimensions of users' enjoyment. Indeed, Foster et al. [48] found that an efficient dialogue of the robot and the successful completion of its tasks significantly affected the perceived intelligence and likeability of the robot. Furthermore, Lee et al. [49] found that when the robot failed in a task all its ratings from users decreased compared to when the robot was successful. These previous findings and the trend of our present results give us motivation to further investigate the possible impact of a QA social robot failure over users' enjoyment and satisfaction.

Here, we also report that we conducted some analyses comparing the question-asking behaviour by gender. However, we did not find any interesting trend or any significant effect worth a mention in the results of this work. Similarly, we also calculated the average interaction time for the participants. However, we could not find any interesting results to report. For example, the average time of interaction for the total number of participants (avg = 123.47 sec, min = 49 sec, max = 399 sec, SD = 74.3 sec) was almost the same as the average time for the users of Robot$_{exhibit}$ (avg = 126.27 sec, min = 49 sec, max = 399 sec, SD = 80.1 sec) and the users of Robot$_{tour}$ (avg = 117.0 sec, min = 50 sec, max = 286 sec, SD = 58.3 sec).

Our findings confirm that, conversely to that expected from observation of human–human interaction, the physical, social presence and situated context of the robot leads to a different question-asking behaviour of users, hence reinforcing the reasoning of Cruz-Sandoval [34], namely that embodiment, degree of anthropomorphism and non-verbal language affect the corpus used in conversational robot systems. Moreover, a similar conclusion was given by Lee et al. [7] that the social norms found in human–human interactions are not always followed during human–robot interactions. From a perspective of improving human–robot interaction, the implications of these findings are that the addition of non-context-specific question-answering ability may enable a more robust, successful and enjoyable service, therefore enhancing the overall user's experience.

As mentioned before, it is infeasible to list all sorts of questions a user can ask. However, by conducting multiple in-the-wild studies, like ours, the types of questions users ask in a particular social setting (e.g. a public one-day event like Open Day) can be modelled. In addition, if the scope of study is limited to a specific domain (e.g. IT), then arguably the performance measures (like, the accuracy of the QA system and user satisfaction) can be improved.

It is worth mentioning, although no formal method was followed, our robots were given a personality. It was done by adopting a consistent tone while adding answers to the questions in the robots' database. We acknowledged that as per [50] people interact with technology in a social manner, and they may unconsciously assign a personality to the robot. Therefore, being a university representative, it was necessary to cater to how the robot's personality was perceived.

8 Conclusion

This study has given an insight into how human–robot interactions for QA social robots may be improved. We found users have least interest in asking context-specific questions to a QA social robot situated in a public event, like a university's open day. Instead, users mostly ask questions which are robot-related or not related to the event. We also discovered adding non-context-specific questions increases the success of a QA system.

Our study presents some limitations. Firstly, our study is specifically for humanoid social robots. Another robot embodiment may lead to a different question-asking behaviour. Secondly, as may be the case for HRI user studies, it is emphasised that this finding is a "snapshot in time" and may not hold once social robots are commonplace and successfully performing in society. Finally, we cannot exclude that interacting users influenced the interaction of users currently queuing and looking at users interacting with the robot. However, this is what expected in real in-the-wild social robots applications deployed in public spaces.

Our future work will include testing different robot personalities appropriate to the role of the robot and including comprehensive measurement of user satisfaction and enjoyment. Furthermore, running this study in different environments and framing conditions may assist in producing results that can be generalised to more application domains.

Acknowledgements This research is supported by an ARC Discovery Project and the Australian Government Research Training Program Scholarship.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

References

1. Paepcke S, Takayama L (2010) Judging a bot by its cover: An experiment on expectation setting for personal robots. In: 2010 5th ACM/IEEE international conference on human-robot interaction (HRI), IEEE, pp 45–52
2. De Graaf M, Ben Allouch S (2014) Expectation setting and personality attribution in HRI. In: Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction, ACM, pp 144–145
3. Goodrich MA, Schultz AC et al (2008) Human-robot interaction: a survey. Found Trends Hum-Comput Interact 1(3):203–275
4. Sheridan TB (2016) Human-robot interaction: status and challenges. Hum Factors 58(4):525–532
5. Leite I, Martinho C, Paiva A (2013) Social robots for long-term interaction: A survey. Int J Soc Robot 5(2):291–308
6. Lee MK, Makatchev M (2009) How do people talk with a robot? An analysis of human-robot dialogues in the real world. In: CHI'09 extended abstracts on human factors in computing systems, ACM, pp 3769–3774
7. Lee MK, Kiesler S, Forlizzi J (2010) Receptionist or information kiosk: How do people talk with a robot? In: Proceedings of the 2010 ACM conference on Computer supported cooperative work, ACM, pp 31–40
8. Jokinen K (2018) Dialogue models for socially intelligent robots. In: International conference on social robotics, Springer, pp 127–138
9. Wainer J, Feil-Seifer DJ, Shell DA, Mataric MJ (2007) Embodiment and human-robot interaction: A task-based perspective. In: The 16th IEEE international symposium on robot and human interactive communication, 2007. RO-MAN 2007, IEEE, pp 872–877
10. Dautenhahn K, Ogden B, Quick T (2002) From embodied to socially embedded agents-implications for interaction-aware robots. Cogn Syst Res 3(3):397–428
11. Pfeifer R, Lungarella M, Iida F (2007) Self-organization, embodiment, and biologically inspired robotics. Science 318(5853):1088–1093
12. Ziemke T (2003) What's that thing called embodiment? In: Proceedings of the annual meeting of the cognitive science society, vol 25
13. Li J (2015) The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int J Hum-Comput Stud 77:23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001
14. Zhao S (2003) Toward a taxonomy of copresence. Presence Teleoper Virtual Environ 12(5):445–455
15. Tonkin M, Vitale J, Ojha S, Clark J, Pfeiffer S, Judge W, Wang X, Williams MA (2017) Embodiment, privacy and social robots: May I remember you? In: Proceedings of the 9th international conference on social robotics (ICSR 2017), Tsukuba, Japan, November 22–24, 2017, Springer International Publishing, pp 506–515, 10.1007/978-3-319-70022-950, https://doi.org/10.1007/978-3-319-70022-9_50
16. Kanda T, Sato R, Saiwaki N, Ishiguro H (2007) A two-month field trial in an elementary school for long-term human-robot interaction. IEEE Trans Rob 23(5):962–971
17. Sung J, Christensen HI, Grinter RE (2009) Robots in the wild: Understanding long-term use. In: Proceedings of the 4th ACM/IEEE international conference on human robot interaction, ACM, pp 45–52
18. Huttenrauch H, Eklundh KS (2002) Fetch-and-carry with CERO: observations from a long-term user study with a service robot. In: Proceedings. 11th IEEE international workshop on robot and human interactive communication, IEEE, pp 158–163
19. Pacchierotti E, Christensen HI, Jensfelt P (2006) Design of an office-guide robot for social interaction studies. In: 2006 IEEE/RSJ international conference on intelligent robots and systems, IEEE, pp 4965–4970
20. Nisimura R, Nishihara Y, Tsurumi R, Lee A, Saruwatari H, Shikano K (2003) Takemaru-kun: Speech-oriented information system for real world research platform. In: Proceedings international workshop on language understanding and agents for real world interaction
21. Moldovan D, Paşca M, Harabagiu S, Surdeanu M (2003) Performance issues and error analysis in an open-domain question answering system. ACM Trans Inf Syst (TOIS) 21(2):133–154
22. Höffner K, Walter S, Marx E, Usbeck R, Lehmann J, Ngonga Ngomo AC (2017) Survey on challenges of question answering in the semantic web. Semantic Web 8(6):895–920
23. Nuccio C, Augello A, Gaglio S, Pilato G (2018) Interaction capabilities of a robotic receptionist. In: International conference on intelligent interactive multimedia systems and services, Springer, pp 171–180
24. Yu L, Hermann KM, Blunsom P, Pulman S (2014) Deep learning for answer sentence selection. arXiv preprint arXiv:1412.1632
25. Kumar A, Irsoy O, Ondruska P, Iyyer M, Bradbury J, Gulrajani I, Zhong V, Paulus R, Socher R (2016) Ask me anything: Dynamic memory networks for natural language processing. In: International conference on machine learning, pp 1378–1387
26. Kim KM, Heo MO, Choi SH, Zhang BT (2017) Deepstory: Video story qa by deep embedded memory networks. arXiv preprint arXiv:1707.00836
27. Xu H, Saenko K (2016) Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European conference on computer vision, Springer, pp 451–466
28. Purington A, Taft JG, Sannon S, Bazarova NN, Taylor SH (2017) Alexa is my new BFF: Social roles, user satisfaction, and person-

ification of the amazon echo. In: Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems, ACM, pp 2853–2859

29. López G, Quesada L, Guerrero LA (2017) Alexa vs. Siri vs. Cortana vs. Google Assistant: A comparison of speech-based natural user interfaces. In: International conference on applied human factors and ergonomics, Springer, pp 241–250
30. Kopp S, Gesellensetter L, Krämer NC, Wachsmuth I (2005) A conversational agent as museum guide–design and evaluation of a real-world application. In: International workshop on intelligent virtual agents, Springer, pp 329–343
31. Luger E, Sellen A (2016) Like having a really bad PA: The gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI conference on human factors in computing systems, ACM, pp 5286–5297
32. Sugiyama H, Meguro T, Higashinaka R (2017) Evaluation of question-answering system about conversational agent's personality. In: Dialogues with social robots, Springer, pp 183–194
33. Cantrell R, Scheutz M, Schermerhorn P, Wu X (2010) Robust spoken instruction understanding for HRI. In: Proceedings of the 5th ACM/IEEE international conference on human-robot interaction, IEEE Press, pp 275–282
34. Cruz-Sandoval D, Eyssel F, Favela J, Sandoval EB (2017) Towards a conversational corpus for human-robot conversations. In: Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction, ACM, pp 99–100
35. Ben-Youssef A, Clavel C, Essid S, Bilac M, Chamoux M, Lim A (2017) UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions. In: Proceedings of the 19th ACM international conference on multimodal interaction, ACM, New York, NY, USA, ICMI 2017, pp 464–472, 10.1145/3136755.3136814, http://doi.acm.org/10.1145/3136755.3136814
36. Bohus D, Saw CW, Horvitz E (2014) Directions robot: In-the-wild experiences and lessons learned. In: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, international foundation for autonomous agents and multiagent systems, pp 637–644
37. Guo S, Lenchner J, Connell J, Dholakia M, Muta H (2017) Conversational bootstrapping and other tricks of a concierge robot. In: Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction, ACM, pp 73–81
38. Iwamura Y, Shiomi M, Kanda T, Ishiguro H, Hagita N (2011) Do elderly people prefer a conversational humanoid as a shopping assistant partner in supermarkets? In: 2011 6th ACM/IEEE international conference on human-robot interaction (HRI), pp 449–457, 10.1145/1957656.1957816
39. Philips L (2000) The double metaphone search algorithm. C/C++ Users J 18(6):38–43
40. Mutlu B, Forlizzi J (2008) Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In: 2008 3rd ACM/IEEE international conference on human-robot interaction (HRI), IEEE, pp 287–294
41. Robson C, McCartan K (2016) Real world research. Wiley, New York
42. Moreno-Sánchez I, Font-Clos F, Corral Á (2016) Large-scale analysis of Zipf's law in English texts. PloS One 11(1):e0147073
43. Glaser BG, Strauss AL (2017) Discovery of grounded theory: strategies for qualitative research. Routledge, London
44. Thomaz A, Hoffman G, Cakmak M (2016) Computational human-robot interaction. Found Trends Robot 4(2–3):105–223
45. Tonkin M, Vitale J, Herse S, Williams MA, Judge W, Wang X (2018) Design methodology for the UX of HRI: A field study of a commercial social robot at an airport. In: Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction, ACM, pp 407–415
46. Reeves B, Nass CI (1996) The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press, Cambridge
47. Hoffmann L, Krämer NC, Lam-Chi A, Kopp S (2009) Media equation revisited: Do users show polite reactions towards an embodied agent? In: International workshop on intelligent virtual agents, Springer, pp 159–165
48. Foster ME, Gaschler A, Giuliani M, Isard A, Pateraki M, Petrick R (2012) Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In: Proceedings of the 14th ACM international conference on Multimodal interaction, ACM, pp 3–10
49. Lee MK, Kiesler S, Forlizzi J, Srinivasa S, Rybski P (2010) Gracefully mitigating breakdowns in robotic services. In: 2010 5th ACM/IEEE international conference on human-robot interaction (HRI), IEEE, pp 203–210
50. Nass C, Moon Y (2000) Machines and mindlessness: Social responses to computers. J Soc Issues 56(1):81–103