CS6750: Assignment M2

Trevor Sands
tsands6@gatech.edu

Abstract—Once considered pure science fiction, technological advances have brought intelligent personal assistants (IPAs) into reality. These virtual agents are now ubiquitous and an expected part of modern personal computing systems. Despite this, users often face gulfs of execution and evaluation when attempting to gather information via the assistants. This project will explore the question and answer (Q&A) task between users and their IPA interfaces.

1 NEEDFINDING EXECUTION

The needfinding plan from assignment M1 called for three distinct approaches: an online **peer survey**, the act of **participant observation**, and an **analysis of existing interfaces** based on direct user feedback. The following subsections will summarize needfinding results and will discuss how the plans controlled for bias.

1.1 Peer Survey

Online surveys allow for input from many users at little-to-no cost to the researcher. For this reason, an online survey of peers (i.e., fellow students in CS6750) was conducted via Georgia Tech's PeerSurvey[1] system.[2]

1.1.1 *Summarized Results*

This survey consisted of questions related to which IPAs users currently interact with, how interaction is conducted (i.e., interface modality), and the holistic feelings of the user about their IPA interactions (e.g., level of comfort performing a Q&A interaction). The survey consisted of 13 questions, with a mix of free text responses, single- and multi-choice questions, and five-point Likert scale (1932) assessments for agreeableness and frequency of interactions.

25 survey responses were gathered, where 9 of the participants were between the ages of 18-29, 8 between the ages of 30-39, and 7 between the ages of 40-

[1] http://peersurvey.cc.gatech.edu

[2] Survey questions and results are provided in the Appendix.

49; one participant declined to provide their age group as the question was optional. Of the 25 responses, 14 users use multiple IPAs and all but one prefer to interact with their IPA by voice (the one outlier prefers interaction via physical keyboard). 14 participants regularly use more than one IPA. Apple's Siri was the most popular with 18 users, Google's Assistant had 13 users, Amazon's Alexa had 11 users, and Microsoft's Cortana had 1 user; the same outlier participant for the interaction modality question selected 'Other' for their IPA but declined to provide information about their IPA of choice.

The agreeableness questions are summarized in Table 1 and the frequency questions are summarized in Table 2.

Table 1—Text and responses for survey questions pertaining to agreeableness with a given statement.

Question Text	S. Disagree	Disagree	Neutral	Agree	S. Agree
"I feel comfortable asking my IPA questions."	1	2	5	12	5
"My IPA understands what I'm asking it."	1	2	6	14	2
"I am satisfied with the answers my IPA provides."	1	1	7	14	2

Table 2—Text and responses for survey questions pertaining to frequency for a given statement.

Question Text	Never	Rarely	Occasionally	Frequently	V. Frequently
How often do you have Q&A interactions with your IPA(s)?	2	7	10	4	2
How often do you need to repeat or rephrase your questions?	1	6	13	5	0
How often are you surprised by your IPA's responses?	2	9	13	1	0

Users were asked to optionally provide context for the types of questions they ask their IPA. Of the 22 responses, most centered around local and regional contexts (e.g., "What's the weather?", "What's the score of the baseball game?"); other common questions were about general facts (e.g., "Can dogs eat 'X'?"). Users were also given the option to suggest interface improvements; 14 participants elected to provide suggestions. Many of these suggestions were purely technical (e.g., "Better voice recognition."), some pertained to not knowing the full scope of the types of questions users could ask a given IPA, and a couple suggested tighter integration with external systems to provide additional context to the IPA during user activities (e.g., exercise, travel).

1.1.2 *Controlling for Bias*

As identified in assignment M1, the two primary biases for the survey were confirmation (Nickerson, 1998) and recall (Last, 2001) biases. The questions were phrased in a neutral manner and consistent numerical evaluation scales were used to control confirmation bias. Recall bias was controlled for by asking participants to have a short Q&A interaction with their IPA prior to taking the survey, but it's unclear as to how many actually did immediately prior to filling out the survey.

One bias that was uncontrolled was the population of the participants themselves. The needfinding survey plan was written prior to knowing about the Peer Survey tool, and originally the participant pool was intended to be more general. As all participants in this survey are fellow Georgia Tech OMSCS students currently enrolled in CS6750, the responses may be more biased towards technically savvy users who may interact with IPAs more frequently than the general population.

1.2 Participant Observation

As naturalistic observation is not practical for this problem space, participant observation was instead used. I followed a script stepping through IPA interactions, which included the types of questions to ask (i.e., local and general) but specific prompts are not given. This same script is provided to others in the third needfinding exercise. The Q&A interaction took place in a home office environment, and my full attention was given to the Q&A interaction.[3]

[3] Interaction script and results are provided in the Appendix.

1.2.1 *Summarized Results*

Prior to performing the interaction, I noted some background information about me as a potential user: I am comfortable performing Q&A interactions with Google's Assistant via a voice interface, and I consider myself adept at performing these interactions. One local question was asked pertaining to when sunset will be for my current location (8:29 PM on June 12). A general question was asked regarding the current price of Bitcoin in USD ($35,983.30 as of the time of this writing)[4] A translation was requested for the Spanish word for "airplane" ("avión"). Finally, I requested that my IPA provide me with an interesting fact.[5]

I found that the IPA understood my questions the first time they were asked. The interface displayed how it interpreted my spoken words, which I found helpful when verifying functionality. I did not need to repeat or rephrase any questions as they were interpreted correctly when first asked, which boosted my confidence in the IPA. While the gulf of evaluation was effectively bridged, I did note that only the "interesting fact" answer was given as auditory feedback in addition to visual feedback in the form of words on the screen; all other questions were presented as visual feedback on screen without accompanying audio. For the context of my interactions (quiet environment, no other cognitive load) this was acceptable, but in cases where one's full attention is not given to the task at hand some information may be lost, necessitating a repeat of the question to the IPA.

1.2.2 *Controlling for Bias*

Confirmation bias is the most significant risk for participant observation, followed by any personal biases I might hold. Confirmation bias can be controlled in the same manner as the survey: writing the interaction script with neutral language and using a consistent scale to rank user agreeableness or frequency. However, I only interacted with Google's Assistant for this evaluation and thus did not get a full picture of other existing IPAs. This may be insufficient for getting a full understanding of the interface designs for different, commonly used IPAs. Furthermore, I conducted this interaction *after* I had received half of the participation script responses, which may have influenced my thinking.

[4] I wish I had listened to my friends about setting up a mining rig in 2010, I could've retired by now.

[5] Today I learned: tsunamis can travel over 500 miles per hour! I'm glad I live inland.

1.3 Evaluation of Existing Interfaces

A small number of participants were asked to follow an interaction script and ask the IPA(s) available to them a series of questions for the purposes of evaluating existing interfaces. Participants follow the same script I used when performing a Q&A task and performed the interactions under similar contexts (at home in a quiet space, full attention to the task).[6]

1.3.1 *Summarized Results*

The interaction script was provided to three individuals known to me[7] as having access to, and having some experience using, an existing IPA. The gathered data is narrower than the peer survey and is meant to test the interfaces by asking a few distinct types of questions; as such, this data is meant to be supplementary to the peer survey.

The three participants covered the three most popular existing IPAs: Apple's Siri (Participant 3), Amazon's Alexa (Participant 1), and Google's Assistant (Participants 2 and 3). Participants 1 and 3 consider themselves novices at performing IPA Q&A interactions; participant 2 considers themselves an expert at performing IPA Q&A interactions. Participants 1, 2, and 3 ranked themselves as "somewhat comfortable (3)", "very comfortable (5)", and "fairly comfortable (4)", respectively. All participants used voice interactions with their IPAs. Two of the participants (1 and 2) did not experience any difficulty in their Q&A interactions nor did they need to repeat any questions, while participant 3 needed to ask the second question ("What is pi to 10 decimal places?") to both Assistant and Siri multiple times before giving up without a satisfactory answer.

When asked about potential features or enhancements:

- Participant 1 found simultaneous visual (text on screen) and auditory (spoken reply) feedback was helpful to evaluating the interaction.
- Participant 2 had no comment about potential features or enhancements.
- Participant 3 noted that both IPA interactions could have been improved by seeking clarification or re-phrasing the question before attempting to answer. They also expressed some disappointment, "Siri should have known more."[8]

[6] Interaction script and results are provided in the Appendix.

[7] These participants are *not* peers from CS6750 nor are they enrolled in OMSCS.

[8] Participant 3 primarily uses Apple products in their day-to-day life.

1.3.2 *Controlling for Bias*

As previously mentioned, confirmation bias was controlled by careful phrasing of the script interaction prompts. However, unlike the other needfinding exercises, this exercise added the potential for observer bias (Mahtani et al., 2018). Observer bias was controlled for participants 1 and 3 by administering the interaction script asynchronously. However, observer bias may have had an effect on participant 2, as I was present while they were performing the interaction (perhaps influencing them to say they are an expert user and are "very comfortable" with the chosen IPA). Furthermore, this limited pool of participants was pulled from my close contacts, which may have unintentionally influenced their responses or willingness to participate.

2 THE DATA INVENTORY FOR ESTABLISHING REQUIREMENTS

To define requirements, the essential questions of the data inventory, discussed in assignment M1, must be answered:

- Who and where are the users?
- What is the context of the task?
- What are the users' goals?
- What do the users need?
- What are the users' tasks and subtasks?

2.1 Answering the Data Inventory

From the survey population, the *who* of the users is implicit: individuals who are technically adept and span a wide age range. This is expected based on the appeal of the OMSCS program to working professionals. A more general view of the population of IPA users would surely include others who are somewhat comfortable working with intelligent technology, despite possibly being less technically savvy. The *where* of the users was overlooked when formulating the questions for the survey, and in the case of participant observation and third-party evaluation of existing interfaces through casual use the context was specified to be in a home environment with no other cognitive load.

The potential *contexts* for the Q&A task were not asked directly, but can be surmised from the survey data by the types of questions users generally ask: directions to somewhere, local information (e.g., weather, transit times), unit conversions, etc. These types of interactions can occur when a user is at rest, planning for near-future actions, or actively in motion trying to get somewhere.

Based on the needfinding activities and the types of questions users generally ask, users' *goals* are generally those of fact-finding, where additional information may inform future decisions (e.g., if I find out what the weather will be later, I'll know if I need to bring an umbrella). In some cases, users indicated that they ask questions based on pure curiosity (e.g., celebrity gossip). In all cases, the users have a *need* for information, most of which is usually accessed through a web browser. The IPA acts as an alternative pathway to the desired information.

The users' *tasks* vary with their goals. In the survey responses that discussed navigation and information about a certain location (e.g., hours of operation), the users' task is to either get somewhere they are not currently. In these cases, subtasks may include determining by what mode they should travel (car, public transit, walking, biking, etc.) In cases where users needed factual information, their task was entirely to be informed; in these cases, the greater context and potential follow-on tasks are not well-defined by the needfinding activities.

2.2 Defining Requirements

Requirements for the IPA Q&A interaction are specified in terms of functionality (what the interface can actually do), usability (how interactions must work), learnability (how quickly can a user understand the interface) and accessibility (who can use the interface); external requirements are driven by needs for compatibility (what devices can the interface run on) and compliance (protection of user privacy). Based on the data inventory and these categories of requirements, the requirements are that the interface must . . .

1. . . . resolve natural language into searchable queries (*Functionality*).
2. . . . infer information where there is ambiguity in the input (*Functionality*).
3. . . . seek additional input where ambiguity cannot be resolved (*Functionality*).
4. . . . provide a hands-free means to ask a question (*Usability*).
5. . . . provide visual and auditory feedback to confirm the question (*Usability*).
6. . . . be intuitive for novice and first-time users (*Learnability*).
7. . . . allow for input by text or spoken natural language (*Accessibility*).
8. . . . understand/respond to the user in their chosen language (*Accessibility*).
9. . . . run on general-use, personal computing devices (*Compatibility*).
10. . . . not reveal personal or identifiable information about the user (*Compliance*).

2.2.1 *Metrics for Evaluation*

As requirements must be measurable, the following metrics will be used to evaluate the quality of a design:

1. Failure rate for natural language interactions (i.e., how often the query failed because of the input it was given).
2. Failure rate for natural language interactions with ambiguity (due to missing words, noisy environment, etc.)
3. True/False if the interface sought additional information.
4. True/False if the interface allows for hands-free queries.
5. Tally of the number and type of modalities used to communicate to the user.
6. Tally of the number of interactions a novice user must have to understand how to ask a question to the interface.
7. True/False if the user allows for spoken and text-based natural language interaction.
8. True/False if interface can parse the user's chosen language.
9. Tally of the number of typical devices (desktop, laptop, tablet, smartphone, smartwatch) the interface can run on.
10. Tally of the number of personal info metadata the interface needs about the user to perform personalized Q&A.

3 FURTHER NEEDFINDING

As discussed in the above section, there is a need for continued needfinding to refine design requirements and better meet the users' actual needs. These follow-on needfinding activities will center around the data inventory questions about the location and context for a user (i.e., under what circumstances is the user interacting with the interface). In addition to this, an IPA interaction will be improved by understanding the user's high-level goals are (i.e., why is the user asking what they're asking). Information about the users' goals will also inform what tasks the user is performing, both with the IPA and with the environment or other individuals.

These additional questions can be answered by a follow-on survey of typical users with more targeted questions. Other sources that can be considered for further needfinding include analysis of product reviews and existing data logs, if such logs are made publicly available.

4 REFERENCES

[1] Last, John M (2001). *A dictionary of epidemiology*. Oxford university press, p. 153.

[2] Likert, Rensis (1932). "A technique for the measurement of attitudes". In: *Archives of psychology*.

[3] Mahtani, Kamal, Spencer, Elizabeth A, Brassey, Jon, and Heneghan, Carl (2018). "Catalogue of bias: observer bias". In: *BMJ evidence-based medicine* 23.1, p. 23.

[4] Nickerson, Raymond S (1998). "Confirmation bias: A ubiquitous phenomenon in many guises". In: *Review of general psychology* 2.2, pp. 175–220.

5 APPENDICES

The following links will expire on August 11, 2021 per Georgia Tech's OneDrive policy. If you would like a copy of the results after this date, please email me.

5.1 Peer Survey

The raw JSON output of the peer survey questions and responses is accessible to Georgia Tech students and faculty via this OneDrive hyperlink.

5.2 Interaction Script

A PDF of the interaction script and results for three participants is accessible to Georgia Tech students and faculty via this OneDrive hyperlink.