Technical Disclosure Commons

Defensive Publications Series

June 11, 2019

Quality survey integrated into a virtual assistant

Xiance Si

Justin Denney

Huayi Li

Quality survey integrated into a virtual assistant

ABSTRACT

This disclosure describes techniques for a virtual assistant to conduct surveys for the quality of responses provided by the virtual assistant to user queries and commands. For screen-free devices, e.g., smart speakers, the surveys are conducted via voice. For devices with a screen, e.g., smartphones, a screen overlay including a survey is displayed, along with a menu of responses. For queries that do not have a clear termination point, surveys are provided via notifications. Surveys are presented in a throttled manner such that a typical user does not get more than one survey question within a certain period. User responses to surveys are utilized to generate a dashboard that provides a visualization of quality ratings and enables the prioritization of features that require optimization.

KEYWORDS

- Virtual assistant
- Query response
- Survey
- User feedback
- Smart speaker
- Smart display

BACKGROUND

The quality of virtual assistant responses, when measured in a continuous, statistically accurate manner, provides developers of the virtual assistant with insight into user satisfaction. Such a quality metric is presently unavailable for virtual assistant software that works across multiple device types such as smartphones, smart speakers, etc. and user interfaces, and offer a

broad feature set across languages. Further, many virtual assistant applications also add new features which are not automatically covered in quality metrics. Also unavailable is a unified view that enables the formulation of a strategy to identify features that result in a lower quality user experience.

For web search queries, human raters measure the quality of a response provided to a single query as input. This technique is ineffective for virtual assistant applications, since the final response to a voice query may be provided after a series of dialog turns rather than a single user query. Human raters find it difficult to understand user context when viewed within logs and are therefore unable to perform realistic assessment regarding whether the responses provided by a virtual assistant application constitute success.

Consider an example where a user instructs a virtual assistant to "set a reminder to pick up laundry at 5 PM." The virtual assistant can possibly receive the instruction in error, e.g., " … at 6 PM," or " … pick up <*unintelligible*> at 5 PM." The accuracy of speech recognition and natural language understanding with which the virtual assistant receives the instruction is best assessed by the user, and not by a post-hoc human rater. In addition, quality ratings by human raters is not scalable across a large feature set, across a large and evolving set of device types, and across several languages.

Other techniques such as yes-no or thumbs-up/down questions provided to users of virtual assistant applications have low response rates. These techniques also suffer from bias, e.g., with responses coming in from only a limited set of highly motivated users, which can make the data statistically invalid.

DESCRIPTION

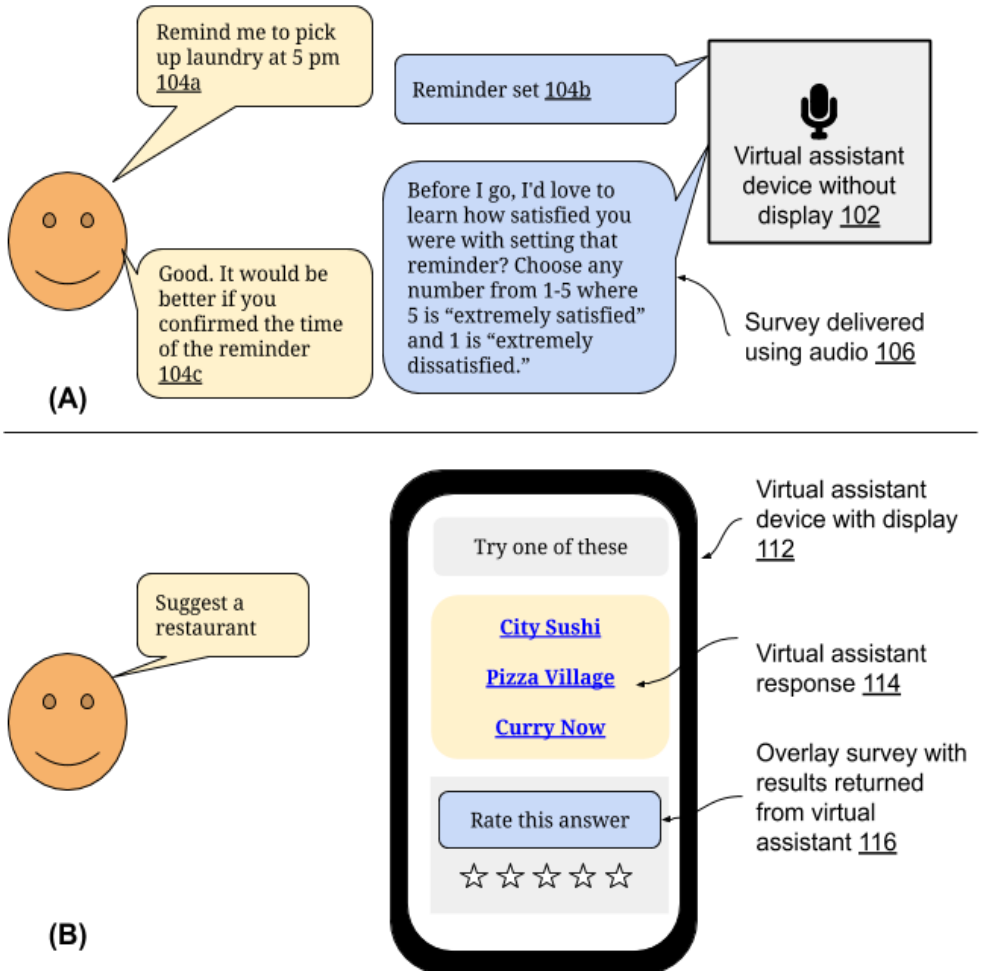

Fig. 1: Quality survey integrated into a virtual assistant (A) Survey on a screen-free device (B) Survey on a device with a screen

Fig. 1(A) illustrates an example of a survey provided by a virtual assistant (102) on the quality of the response provided on a device that lacks a screen. The virtual assistant provides a response to a command, query, etc., or otherwise concludes a conversation (104a-c). The virtual assistant requests the user to respond to a survey on the quality of the response (106). The survey is delivered via voice.

Fig. 1(B) illustrates an example of a virtual assistant requesting a survey on the quality of the response provided on a device with a display (112). The virtual assistant provides a response

(114) to a command, query, etc. The virtual assistant displays a survey (116) to obtain user feedback regarding the quality of the responses provided by the virtual assistant. The survey can be a screen overlay returned with the response. The survey can be in the form of a rating system, a yes-no question, short answer, emojis representing user satisfaction, etc. The survey can be designed such that it is consistent across voice-only devices and devices with displays. Further, if the rating is low, a follow-up question, e.g., "what went wrong?" can be delivered to the user to obtain additional user feedback.

For queries that do not have a clear termination point ("play jazz music") or for commands that take the user outside of the virtual assistant application ("open a map"), surveys are provided via notifications. The notification can appear even when the virtual assistant application is not open. The notification can include content such as "Help improve the virtual assistant. How satisfied were you with the choice of music?" When the user opens the notification, details of the interaction such as a transcription of the user query, the response provided by the virtual assistant, date and time of interaction, device type, etc. are provided for the purpose of easy and accurate recall of the interaction.

To ensure a statistically significant sample for the broad set of features provided by a virtual assistant, the probability with which a survey is triggered is set separately for different features on different device types. For example, popular features of a virtual assistant such as the weather forecast are set to a low probability of triggering a survey. Less frequently used features such as creating calendar events are set to a higher probability of triggering a survey. Surveys are presented in a throttled manner such that any particular user does not get more than one survey question within a certain period, e.g., thirty days, across device types owned by the user.

Survey responses are combined into a single database and displayed via a dashboard that visualizes the average quality ratings across features and devices. A traffic-weighted headroom analysis that takes into account variables such as combination of average rating score, the traffic of features across devices, etc. which allows prioritization of features that require quality optimization.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's social network, social actions or activities, profession, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques for a virtual assistant to conduct surveys for the quality of responses provided by the virtual assistant to user queries and commands. For screen-free devices, e.g., smart speakers, the surveys are conducted via voice. For devices with a screen, e.g., smartphones, a screen overlay including a survey is displayed, along with a menu of responses. For queries that do not have a clear termination point, surveys are provided via notifications. Surveys are presented in a throttled manner such that a typical user does not get

more than one survey question within a certain period. User responses to surveys are utilized to generate a dashboard that provides a visualization of quality ratings and enables the prioritization of features that require optimization.