Adaptive Summarization of Digital Assistant Responses to User Queries

Matthew Sharifi
Victor Carbune

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.
Adaptive Summarization of Digital Assistant Responses to User Queries

ABSTRACT

This disclosure describes techniques for performing adaptive summarization of responses provided by a digital assistant to a user’s queries based on explicit and/or implicit feedback from the user that is indicative of their engagement with given responses. The feedback is determined based on user-permitted data from device sensors. If the user permits, past reactions of the user to output responses are used to train machine learning models or heuristics/rules that are used to determine the length of a query response appropriate for the user. Described features enable a digital assistant to automatically match the response length (duration) to the user’s query type, context, and preferences. This enables the digital assistant to provide an engaging interaction experience, since the user is unlikely to need to interrupt responses or disengage from the digital assistant midway through a response.

KEYWORDS

- Response length
- Response duration
- Response summarization
- Text summarization
- Query response
- Gaze detection
- Attention detection
- Human sensing
- Virtual assistant
- Digital assistant

BACKGROUND

Users frequently interact with digital assistants via smart devices such as phones, watches, and smart speakers. For example, in response to a spoken command, these digital
assistants perform various tasks and find answers to questions posed by the users through natural and conversational interaction.

While voice-based interfaces of digital assistants are convenient to use, it can be challenging for a digital assistant to determine the appropriate length of a response to a user query based on whether the user is best served by a detailed or a summarized response for a given query. This can sometimes result in an overly short or an excessively long response which can lead to the user losing engagement with or attention to the digital assistant, or having to interrupt the response midway through the output.

**DESCRIPTION**

This disclosure describes features that perform adaptive summarization of digital assistant responses based on explicit and/or implicit feedback (obtained with user permission) from users that is indicative of their engagement with a response provided by the digital assistant. The digital assistant can adapt the length of the response by performing response summarization based on the user’s current context and the type of query, as indicated by past interactions with digital assistant responses. The described techniques are implemented with specific user permission to access a user’s data, e.g., speech data from the user, images/video of the user, user context information, etc. Users are provided with options to grant permissions to such data and/or to disable the features entirely. The user can enable or disable techniques discussed herein for particular locations, time periods, or for other conditions.

**Example of use**

For example, in response to the user query “Who is Barack Obama?” responses of different lengths may be as follows:

- **Short:** “Barack Obama is a former president of the United States”
- **Medium**: “Barack Obama is a former president of the United States. He served from 2009 to 2017”

- **Long**: “Barack Obama is a former president of the United States. He served from 2009 to 2017. A member of the Democratic Party, Obama was the first African-American president.”

The appropriate length for a particular query may be determined using techniques described herein. While the above example shows three different lengths, responses of any duration or length can be provided as appropriate. Further, as described below with reference to Fig. 1, the length of the provided response can be adjusted dynamically.

**Example Operational Implementation**

![Diagram](image)

**Fig. 1**: Digital assistant that performs adaptive summarization of assistant responses
Fig. 1 shows an example of operational implementation of techniques described in this disclosure. A user (102) issues a spoken query (104) that is received by a device (106) that provides a digital assistant. The query can be a question seeking an answer about the user’s schedule, facts, device characteristics, or any other subject. The device can be a user device, e.g., smartphone, wearable device, a smart speaker, or other device.

The query is provided as input to a digital assistant (108) that can be an application, a module in an operating system, etc. With user permission, the query is processed by a stack (110) of the digital assistant. For example, the stack can include an Automatic Speech Recognition (ASR) system that includes trained machine learning models and/or other components to perform speech recognition. The assistant stack can also include natural-language understanding (NLU) components to process the query.

One or more candidate responses to the query are generated by the stack. For example, a candidate response can be generated by extracting a snippet from a data source (112), where the snippet relates to the query. The data source can include one or more of a structured knowledge graph, one or more web pages, or other information sources. A large amount of information that includes different pieces can be included in a candidate response, such that the candidate response may be very long. By default, response generation starts by selecting longer response snippets, since such snippets are easier to reduce in size based on indicators detected for the user.

A candidate response may be adjusted to an appropriate summarized response length (e.g., snippet) for the user based on the use of response length models (114). The response length models can be statistical, rule-based models or can be machine learning models. With user permission, these models predict an appropriate length of a response based on a current context of the user (e.g., user activity such as driving or cooking, location or environment of user, etc.) as
indicated in audio and/or visual data (116) detected by sensing devices (118), such as a camera and/or microphone of the device and/or as indicated in other data, e.g., location data from a GPS sensor of the device, motion data from accelerometer or gyroscope of device, etc., accessed with user permission.

In addition, the appropriate response length can be based on the type of the query that was received from the user, e.g., a factual question, a request for user data, etc. The response length models can be trained to determine an appropriate length for a response based on various user contexts and types of queries, using a training data corpus obtained from volunteers and/or the specific user using the device, as described in greater detail below. When training for the specific user is performed, user data such as past queries and responses as well as past context and user reaction to responses is utilized for training purposes only if the user permits such training. The training is performed locally on-device such that the user data does not leave the user device.

After summarizing the obtained information, the response having an appropriate length is provided to the stack. The summarized response can be generated using any of various techniques. For example, if a shorter response is determined to be appropriate, a long candidate response can be shortened. In some examples, standard extractive text summarization techniques can be used to create a snippet that acts as a shorter response. Alternatively, the number of pieces of information that are stitched together in the summarized response can be limited based on the appropriate response length. Alternatively a snippet of variable and appropriate length can be taken from the data source. If a long response is found to be appropriate, a too-short generated response can be lengthened using similar techniques.
The response (120) having the appropriate length is output to the user, e.g., by a text-to-speech engine (122) that outputs the response. The response can be output as audio to the user from speakers of the device. In some cases, the response can be additionally or alternatively provided in visual form, e.g., displayed as text on a screen or other display device.

**Training Response Length Models and Using the Models to Adjust Response Delivery**

To train the response length models, user reactions to responses are, with user permission, detected during the output of responses. A number of different indicators related to the user reactions can be examined. The indicators are used to determine whether an output response is of an inappropriate length for the user, e.g., too long or too short. These indicators can include the following, which may be used to indicate that the response is too long:

- **Human sensing**: detecting that the user has moved away from the device, e.g., from a smart speaker or other device
- **Audio/visual**: detecting the user has started an activity that indicate distraction, such as engaging in conversation with another person
- **User gaze**: detecting that the user is looking away from the device that is providing the response; this is particularly useful for visual responses such as photos or text.
- **Interruptions**: detecting that the user has interrupted the query, e.g., by saying “stop” or “thanks.”

The response length models are trained based on the indicators detected during output of responses, if users permit such training. The models can be global models and/or user-specific models (e.g., trained locally with data that is associated to the specific user). The response length models are trained to predict the appropriate length for a response based on a context of the user and based on a type of query. For example, a model can be trained for different user contexts.
(e.g., driving, cooking, at a particular location, etc.), or a single model can be conditioned on a particular user context. Similarly, a model can be trained for multiple types of queries, e.g., query verticals or domains (e.g., longer responses are appropriate for “weather”; shorter responses are preferred for “sports scores”; etc.), or individual models can be trained for different query types.

If the user permits, the response length models can be fine-tuned on an ongoing basis, e.g., updated over time based on additional received queries from the user and based on detected user reactions to output responses. Thus, query responses having lengths more customized to the specific user can be generated over time.

At the time of delivery, it can be detected that the response length is inappropriate, e.g., too long. With user permission, context data of the user for the one or more of the above indicators can be sensed using one or more of the sensing devices such as a device camera, device microphone, location sensor, proximity sensor, etc. With user permission, context data can be examined to detect one or more of the above indicators by using indicator models (124), e.g., that include dedicated machine learning models for each type of user behavior. The indicator models can include deep neural networks trained for a specific task on labeled context data, e.g., to detect a gaze of the user at the camera, detect interaction with another person, detect speech of the user with particular characteristics, etc. The indicator models output whether any of the above indicators have been detected in user behavior while the response is being provided to the user.

Based on one or more of the above indicators being detected, a candidate response can be determined as being too long by the response length models. For example, based on the indicators, disengagement of the user can be found at a particular point in the output of the
response prior to its completion. This may indicate that the response is too long, or may indicate that the user already obtained the information that they were seeking prior to response completion. The former case can be distinguished from the latter case based on multiple detected user reactions for various responses.

Indicators can also be used to detect whether an output response is too short for the user. One example of such an indicator is detecting a follow-up user query, e.g., a user saying, “tell me more” immediately after the completion of response output. Such an indicator may indicate that the response is too short and did not include enough information for the user.

Adaptive query summarization can be performed in a “live” manner. For example, if some of the indicators of user disengagement described above are detected during response output that indicate an overly-long response, the response output can be stopped at the next natural break point of the response instead of the entire response being output.

While Fig. 1 shows a digital assistant implemented on a device, described techniques can be implemented in other configurations as permitted by the user. For example, the functionality of the digital assistant and/or other components may be incorporated within one or more applications or in a device operating system. In another example, if the user permits, one or more of the digital assistant, models, text-to-speech engine, data sources, and/or other components can be executed on a server that is remote from the user device. The various features described herein are implemented only with user permission to access user data that serves as input to the system.

Further to the descriptions above, a user is 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., context information about a user, a user’s speech
input and commands, a user’s responses to information delivered by a virtual assistant, a user’s preferences, a user’s location, a user’s calendar items, etc.), 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. 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 performing adaptive summarization of responses provided by a digital assistant to a user’s queries based on explicit and/or implicit feedback from the user that is indicative of their engagement with given responses. The feedback is determined based on user-permitted data from device sensors. If the user permits, past reactions of the user to output responses are used to train machine learning models or heuristics/rules that are used to determine the length of a query response appropriate for the user. Described features enable a digital assistant to automatically match the response length (duration) to the user’s query type, context, and preferences. This enables the digital assistant to provide an engaging interaction experience, since the user is unlikely to need to interrupt responses or disengage from the digital assistant midway through a response.