February 2021

Improved Contextual Grounding by Combining Multiple Speech Transcription Hypotheses

Manaal Faruqui

Vishal Verma

Aditya Gupta

Improved Contextual Grounding by Combining Multiple Speech Transcription Hypotheses

ABSTRACT

Processing a user’s voice command includes parsing the command to derive the components referred to therein. These identified components or arguments are then mapped to items or objects in the real world in a process known as “grounding.” In some cases, transcription inaccuracies can make it infeasible for a virtual assistant or other application to achieve accurate grounding, thus making it impossible to service the user’s command. This disclosure describes techniques to improve grounding by taking into account the top N highest-likelihood transcriptions for a user’s voice command along with contextual information accessed with the user’s permission. Improved query interpretation can enable a virtual assistant or other application to accurately interpret the command and thereby improve user experience.

KEYWORDS

- Grounding
- Virtual assistant
- Automatic Speech Recognition (ASR)
- Natural Language Understanding (NLU)
- Query interpretation
- Speech transcript
- Transcription error
- Transcription context
- Entity recognition
- Smart speaker
- Smart display
- Smart appliance

BACKGROUND

Users often provide speech commands to a voice-based virtual assistant, e.g., available via a device such as smartphone, smart speaker, etc. instructing the device to perform a task such as setting a timer, starting media playback, etc. The user’s voice input is first transcribed into text using Automatic Speech Recognition (ASR). ASR typically includes generating candidate hypotheses for the user's spoken input and choosing the text matching the hypothesis that has the highest likelihood of a match with the user's input. The transcribed text is then passed to the virtual assistant to fulfil the user's request.

To that end, the transcribed text is processed via a Natural Language Understanding (NLU) model to parse the user's spoken request to derive relevant embedded components, such as the command, the entities, etc. For instance, a request to "cancel the pizza timer" may be parsed to contain the "CancelTimer" command with a "timer" entity labeled "pizza." This process of mapping the components of a user's command to entities is termed as "grounding."

The transcription produced by ASR is not always accurate. In some cases, transcription inaccuracies can make it infeasible for the virtual assistant to service the user's request. For example, if the ASR erroneously transcribes the utterance "pizza" as "pita," the user's request to cancel a previously set "pizza timer" would be inferred as a request to cancel a "pita timer." Since no pita timer exists, the response from the virtual assistant would indicate that the virtual assistant cannot find a pita timer to cancel.

DESCRIPTION

This disclosure describes techniques to achieve better grounding for a user's voice commands. To that end, with specific user permission, the techniques involve taking into account the first N ASR transcription hypotheses with the highest likelihood of matching the user's spoken input to a voice-based virtual assistant when performing query interpretation.

The grounding process is performed by NLU processing of multiple hypotheses to generate a combined list of components, such as commands, entities, etc., contained within these hypotheses. The NLU output based on combining the N highest likely transcription hypotheses is then considered in relation to the relevant information on the user's context, obtained with the user's permission.

For example, consider that a user issues a spoken command "cancel the pizza timer," indicating the desire to cancel a pizza timer set previously. The ASR process may produce the following transcription hypotheses for the user's voice input:

1. "Cancel the pita timer"
2. "Cancel the pizza timer"
3. "Cancel dapita timer"

The hypotheses can be ranked based on the likelihood of match with the user's input as indicated by corresponding ASR confidence scores. NLU processing of the three hypotheses taken together can employ the Hirschberg alignment to identify component tokens that correspond to the same entity in the user's speech.

In the above example, all three hypotheses refer to a command to cancel a timer with a label that could be "pita," "pizza," or "dapita," respectively in order of likelihood. Therefore, the NLU parse results in a command to cancel a timer with a label pita, pizza, or dapita, in that order. Since the context information indicates no timer labeled "pita," the command interpretation continues down the list to look for a "pizza" timer. The command thus results in the user's previously set pizza timer being found and canceled, instead of producing an error message regarding the virtual assistant being unable to find a timer labeled pita.

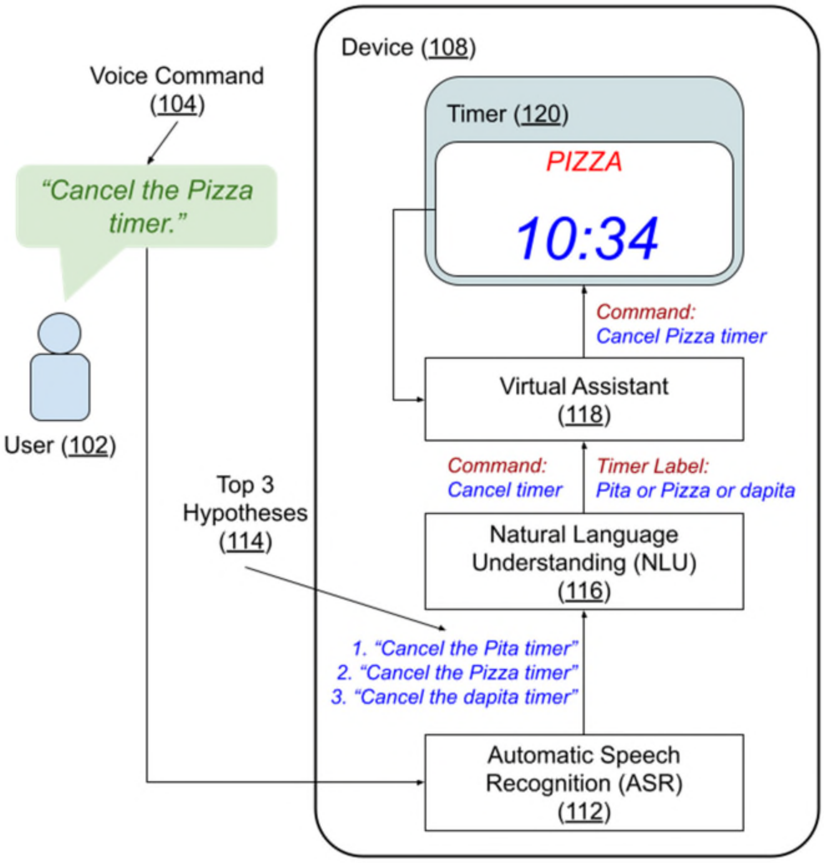
Device (108)
Voice Command (104)
"Cancel the Pizza timer."
User (102)
Timer (120)
PIZZA
10:34
Command: Cancel Pizza timer
Virtual Assistant (118)
Command: Cancel timer
Timer Label: Pita or Pizza or dapita
Top 3 Hypotheses (114)
Natural Language Understanding (NLU) (116)
1. "Cancel the Pita timer"
2. "Cancel the Pizza timer"
3. "Cancel the dapita timer"
Automatic Speech Recognition (ASR) (112)

Fig. 1: Combining top ASR hypotheses to achieve better contextual grounding

Fig. 1 shows an example of operational implementation of the techniques described in this disclosure. A user (102) issues a voice command (104) to a virtual assistant (118) available via a device (108). The command is a request to cancel a previously set timer labeled "pizza" (120). With permission, the user's speech is processed by ASR (112) to generate a text transcription of the voice input.

The top three hypotheses for the transcriptions with the highest likelihood of matching the user's voice input are processed in combination by the NLU module (116) to derive the command ("cancel timer") and parameters (timer labels) contained within them. The virtual assistant (118) is provided the output of the NLU parse along with relevant context information that indicates that a timer labeled pizza is currently running. Based on the NLU parse and the context obtained with user permission, the virtual assistant fulfills the user's request by canceling the pizza timer.

The techniques described above can also handle situations that involve the reverse situation wherein a user's previous command may have been incorrectly transcribed. For instance, a user's request to "set a sale timer for 90 minutes" can result in setting a 90-minute timer labeled "cell" because of erroneous ASR. When the user later asks for the time remaining on the sale timer, the above operation can provide the desired answer (the time remaining on the timer labeled "cell") even though its initial label was not set accurately. Further, while the above examples illustrate different interpretations of a single entity or word ("pizza"), more complex commands may include multiple entities and/or actions and may be interpreted in a similar fashion. For example, the command "cancel the pizza timer and set a reminder to pick up laundry" has multiple actions ("cancel timer" and "set a reminder") each with its own associated entities.

The various automated processes, such as ASR and NLU, involved in the above described operations can use any suitable trained machine learning models. The threshold values and the number of ASR hypotheses N used in the operation of the techniques can be set by the developers and/or specified by the users and/or determined dynamically at runtime.

Implementation of the techniques can improve contextual grounding for voice input without much increase in computation costs or latency. The techniques can be embedded within any service, platform, or application that provides voice based virtual assistant capabilities. The described operation can serve to enhance the user experience (UX) of using a voice-based virtual assistant for various contextual tasks, such as home automation, alarms, timers, etc.

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 voice commands, contextual information, or a user's preferences), 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 to improve grounding by taking into account the top N highest-likelihood transcriptions for a user's voice command. The grounding process is performed by applying natural language understanding to process multiple hypotheses and generating a combined list of components, such as commands, entities, etc., contained within these hypotheses. The NLU output based on combining the N highest likely transcription hypotheses is evaluated in relation to relevant information on the user's context obtained with the user's permission. Improved query interpretation can enable a virtual assistant or other application to accurately interpret the command and thereby improve user experience.

REFERENCES

1. Mengibar, Pedro J. Moreno, Fadi Biadsy, and Diego Melendo Casado. "Evaluating transcriptions with a semantic parser." U.S. Patent 8,868,409, issued October 21, 2014.
2. Hirschberg's algorithm https://en.wikipedia.org/wiki/Hirschberg%27s_algorithm accessed February 15, 2021.
3. Faruqui, Manaal and Christensen, Janara, "Contextual Error Correction in Automatic Speech Recognition", Technical Disclosure Commons, (March 06, 2020) https://www.tdcommons.org/dpubs_series/2989