October 2020

Incorporating Device Context In Natural Language Understanding

Anonymous

Incorporating Device Context In Natural Language Understanding

ABSTRACT

Automatic speech recognition (ASR) models are used to recognize user commands or queries in products such as smartphones, smart speakers/displays, and other products that enable speech interaction. Automatic speech recognition is a complex problem that requires correct processing of the acoustic and semantic signals from the voice input. Natural language understanding (NLU) systems sometimes fail to correctly interpret utterances that are associated with multiple possible intents. Per techniques described herein, device context features such as the identity of the foreground application and other information is utilized to disambiguate intent for a voice query. Incorporating device context as input to NLU models leads to improvement in the ability of the NLU models to correctly interpret utterances with ambiguous intent.

KEYWORDS

- Automatic Speech Recognition (ASR)
- Natural Language Understanding (NLU)
- Speech model
- Device context
- Virtual assistant
- Smart speaker

BACKGROUND

Automatic speech recognition (ASR) models are used to recognize user commands or queries in products such as smartphones, smart speakers/displays, and other products that enable speech interaction. Automatic speech recognition is a complex problem that requires correct processing of the acoustic and semantic signals from the voice input.

ASR typically involves the use of an acoustic model that transforms an input audio signal (e.g., a spoken query or command from a user) into a feature vector and a natural language understanding (NLU) model that uses the features to obtain a meaningful utterance. Accurate language understanding can utilize information from previous conversational history and information about the environment, referred to as context. Some current NLU models use context features, such as the user's contact list, active applications, and/or previous queries.

However, NLU systems sometimes fail to correctly interpret utterances that are associated with multiple possible intents. For example, a simple utterance such as “play Halo” can be classified into different domains such as games or music. Conventional NLU models typically use a set of predefined rules for disambiguation in such situations. For example, if media is prioritized over gaming, “play Halo” is interpreted as a command to play the song “Halo” by Beyoncé.

However, a simple rule-based approach is not generalizable in a satisfactory way across multiple domains. For example, an NLU model that always prioritizes music over other domains may not be able to interpret the command “put on Taylor Swift” as a request to place a call to a contact named Taylor Swift and instead may provide an inappropriate response of starting playback of songs by the artist Taylor Swift. This is especially likely when the context in which the command is received includes a music playback app in the foreground.

DESCRIPTION

This disclosure describes techniques to improve the performance of NLU models by incorporating device context when performing disambiguation of intent across domains. The device context can include, e.g., identity of the foreground application (as a variable), of a homepage that displays search results, etc. With user permission, such information is provided as an input to NLU models to improve the prediction of user intent.

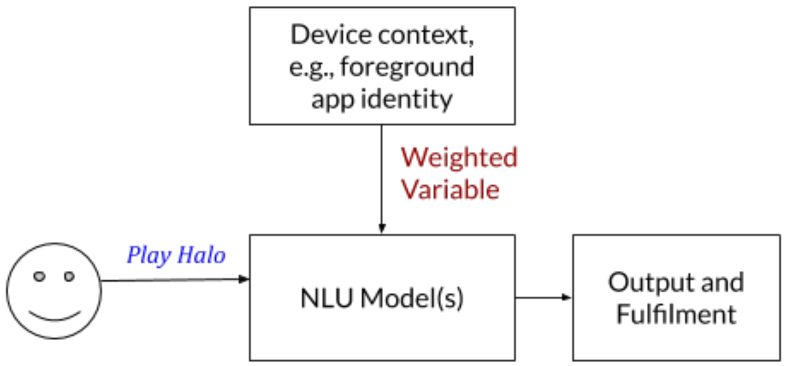

Fig. 1

Fig. 1 illustrates an example process for incorporating device context in NLU models, per techniques of this disclosure. An audio signal (e.g., the spoken user command "Play Halo") is passed to the NLU model(s) for interpretation. The device context variable, which is typically available in the form of an identifier of the foreground app is also passed as an input to the NLU models. The device context variable can be assigned a weight. The value of the weight can be optimized during a model training phase. Further, to avoid overfitting to particular domains (which can cause incorrect query interpretation), a training dataset that includes randomly added device context features for various utterances can be utilized.

In the "play Halo" example shown in Fig. 1, if the device context is "Gaming" (e.g., particular game or game selection UI is in the foreground), input of the device context to the NLU model(s) can lead to the NLU model(s) interpreting the spoken input correctly as a command to play the game Halo. On the other hand, the device context is "music playback", the utterance is correctly interpreted as a command to play the song "Halo" by Beyoncé.

With the device context feature provided as an input, the NLU model can thus disambiguate utterances with ambiguous intent. Some examples of context-based query disambiguation are listed below:

- **"Put on baby shark":** A conventional NLU model that doesn't take into account the device context may incorrectly interpret this command as a media request. However, depending on the context, the request may be to play the video "baby shark" when a video app is in the foreground, a request to play the song "baby shark: when a music playback app is in the foreground, or a request to add an augmented reality mask of a baby shark when the camera is in the foreground, e.g., during a call or when taking pictures. By incorporating the device context as described herein, the NLU model can incorporate the appropriate domain such that the utterance is correctly interpreted.
- **"Pure water":** A conventional NLU model that doesn't take into account the device context may incorrectly interpret this command; however, by incorporating the device context as described herein, the NLU model can incorporate the domain of "music playback" the utterance is correctly interpreted as being a request to play the song "pure water."
- **"Put cat ears on my head":** When this command is received and the device context is that the device camera is on, the NLU model can correctly interpret this as a command to add the mask effect of cat ears to the displayed image that is captured by the camera.

While foreground application is described as an example, device context features can include current URL, search results being displayed, volume level, whether a call is in progress, screen brightness, whether a device camera is on, or other suitable contextual features. Further, such features may be different for different devices, e.g., for a smartphone, the device context may include foreground app (or specific portion of app); for a device that includes a search-based interface, the device context may include current URL; etc. For example, when the utterance is “Chase” a normal interpretation may be to conduct a search for “Chase bank near me.” However, if the device context indicates that the “Chase” app is installed on the smartphone where the query is received, the NLU model can instead interpret the query as a request to “Open Chase App.” A mapping mechanism can be provided to adapt the NLU model to different types of devices.

CONCLUSION

Per techniques described herein, device context features such as the identity of the foreground application and other information is utilized to disambiguate intent for a voice query. Incorporating device context as input to NLU models leads to improvement in the ability of the NLU models to correctly interpret utterances with ambiguous intent.