Technical Disclosure Commons

Defensive Publications Series

December 2019

Recognition of spelled out words in spoken queries

Felix Weissenberger

Victor Cărbune

Bogdan Prisacari

Recognition of spelled out words in spoken queries

ABSTRACT

This disclosure describes techniques to enhance automated speech recognition by enabling automatic recognition of words spelled out by users. Machine learning techniques are utilized to detect explicit user intent to spell out a word as well as detect spelled out words without an explicitly stated user intent. If it is determined that the user is spelling a word, a spelling mode is triggered wherein received letters are concatenated together to form a word. If the user permits, data that includes the user context, audio of the word, audio of the user spelling out the word, and the textual representation of the word are obtained and utilized for training. The trained machine learning model is utilized in subsequent processing of user speech.

KEYWORDS

- Query interpretation
- Speech recognition
- Audio classification model
- Voice command
- Natural language processing (NLP)
- Voice assistant
- Smart speaker
- Spoken query
- Spelled query

BACKGROUND

A user's interaction with a virtual assistant typically involves spoken requests, queries, and commands. Such queries and commands are interpreted using automated speech recognition (ASR) techniques. However, ASR techniques can sometimes fail to recognize spoken words, e.g. names, locations, etc. User context based weighting, e.g. based on contacts associated with the user, is commonly utilized to enhance and improve recognition quality. However, words that are not in the context, e.g. a new or rare contact name that the user wants to input for the first time, or words that are phonetically close to a known word can still pose a challenge to automatic speech recognition.

DESCRIPTION

This disclosure describes techniques to enhance automated speech recognition by enabling automatic recognition of words spelled out by users. Per techniques of this disclosure, a user can spell out words that were not recognized or recognized incorrectly when first spoken by the user and the spelled query is used for recognition. With user permission, the spelled out words are also utilized for training and weighting the automated speech recognition for subsequent use. The techniques can be used in any context, e.g., by a virtual assistant that responds to spoken commands or queries. Fig. 1 illustrates an example of recognition of a spelled out word, per techniques of this disclosure.

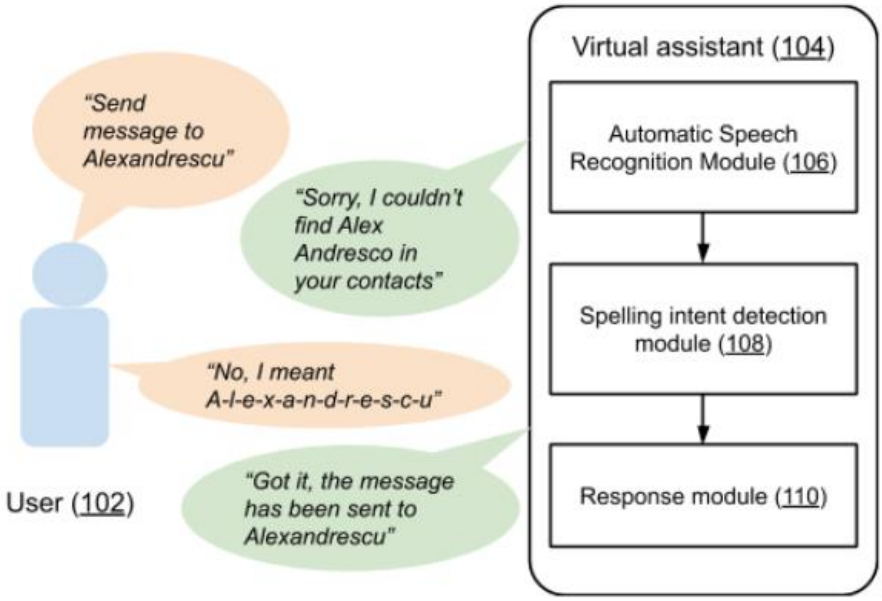

Fig. 1: Interpretation of spelled queries

In this illustrative example, a user (102) intending to send a message to a contact makes a request (*"*Send the message to Alexandrescu") to a virtual assistant (104). The virtual assistant, using an automated speech recognition module (106), interprets the request as a request to send a message to "Alex Andresco" and indicates to the user that it was unable to locate the contact ("Sorry, I couldn't find Alex Andresco in your contacts").

The user proceeds to correct the query by spelling out the word (letter by letter) that was misinterpreted by the virtual assistant ("No, I meant A-l-e-x-a-n-d-r-e-s-c-u"). The virtual assistant detects the user intent to spell out a word via a spelling intent detection module (108) and correctly recognizes the word as intended by the user by concatenation of the letters into the word. The virtual assistant performs the requested action using a response module (110) and provides a confirmation to the user ("Got it, the message has been sent to Alexandrescu").

In this illustrative example, the user explicitly indicates their intention to spell out the incorrectly recognized word by stating "No, I meant" before proceeding to spell out the word.

The spelling intent detection module is trained to detect such explicitly stated user intention (command) to spell out a word.

The spelling intent detection module is also trained to detect user intention to spell out a word without an explicitly stated user intent to spell out the word, e.g., where the user simply states “send a message to A-l-e-x-a-n-d-r-e-s-c-u” in the spoken request or command.

The spelling intention detection module utilizes a machine learning model that indicates whether speech or text input being processed is likely a word being spelled out, and is not a sequence of separated characters. If it is determined that the user is spelling a word, a spelling mode is triggered wherein received letters are concatenated together. The model is also utilized to classify the concatenated word and provide suitable capitalization.

The machine learning model can be a simple 1D convolutional neural network that operates on the received speech or text input and produces a classification output. When the user permits, the model is trained on previously obtained data that includes spelled out words. For example, speech that includes explicitly spelled out words (as in the example described earlier) can be used for model training purposes.

The model can also be used to verify speech that is interpreted by the spelling intent detection module to include explicitly stated user intention to spell out a word.

During operation of the speech recognition module, if the user permits, data that includes the user context, e.g. phrase that included the word, application in use, etc.; spoken version (audio) of the complete word; audio of the user spelling out the word; and the correct textual representation of the word are obtained and utilized for training the machine learning model.

Data pairs that include the audio of the word and the text form can be made available as training data for speech recognition either at a server or on local devices for federated learning. The trained machine learning model is utilized in subsequent processing of user speech to adjust the automatic speech recognition. For example, user pronunciation of certain words can be matched to previously obtained pronunciation by the user, and based on the match, the speech recognition can be weighted to produce the previously spelled-out word as its output, without the user having to spell it out this time, rather than a different matching word. In the example illustrated above, the virtual assistant can assign a higher weight to the probability that the word is “Alexandrescu” than that for “Alex Andresco” in subsequent interactions with the user.

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 to enhance automated speech recognition by enabling automatic recognition of words spelled out by users. Machine learning techniques are utilized to detect explicit user intent to spell out a word as well as detect spelled out words without an explicitly stated user intent. If it is determined that the user is spelling a word, a spelling mode is triggered wherein received letters are concatenated together to form a word. If the user permits, data that includes the user context, audio of the word, audio of the user spelling out the word, and the textual representation of the word are obtained and utilized for training. The trained machine learning model is utilized in subsequent processing of user speech.

REFERENCES

[1] Chung, Grace, Stephanie Seneff, and Chao Wang. "Automatic acquisition of names using speak and spell mode in spoken dialogue systems." In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pp. 32-39. Association for Computational Linguistics, 2003.