Technical Disclosure Commons

Defensive Publications Series

March 2020

Error Correction in Automatic Speech Recognition

Ágoston Weisz

Ragnar Groot Koerkamp

Error Correction in Automatic Speech Recognition

ABSTRACT

This disclosure describes techniques to correct errors in automatic speech recognition, e.g., as performed to recognize spoken queries from a user to a virtual assistant or other application. A machine learning model detects potentially misrecognized n-grams within transcribed text which are then underlined in a user interface. A user can tap on the underlined n-gram, or another portion of the transcribed text to activate a dropdown menu that presents alternatives to the transcribed text. The alternatives can be based on speech hypothesis scores. To correct the error in transcribed text, the user picks an alternative from the dropdown menu, or, in the absence of a suitable alternative, types in the correction. With user permission, the error and corresponding correction are used as training data to improve model performance.

KEYWORDS

- Automatic speech recognition (ASR)
- Voice recognition
- Speech recognition error
- Speech hypothesis scores
- Audio model
- Smart speaker
- Smart display
- Spoken query
- Virtual assistant
- Language model

BACKGROUND

Automatic speech recognition (ASR) is used in a variety of applications, e.g., to recognize spoken queries received by virtual assistants on devices such as smart speakers, smartphones, tablets, etc. Errors in ASR are disruptive to the user-machine interaction. Current techniques to correct ASR errors are cumbersome and often lead to the user reissuing the same query, contributing to user dissatisfaction.

DESCRIPTION

This disclosure describes techniques that enable a user to easily and accurately re-issue utterances misrecognized by an automatic speech recognizer.

Example user interface and workflow

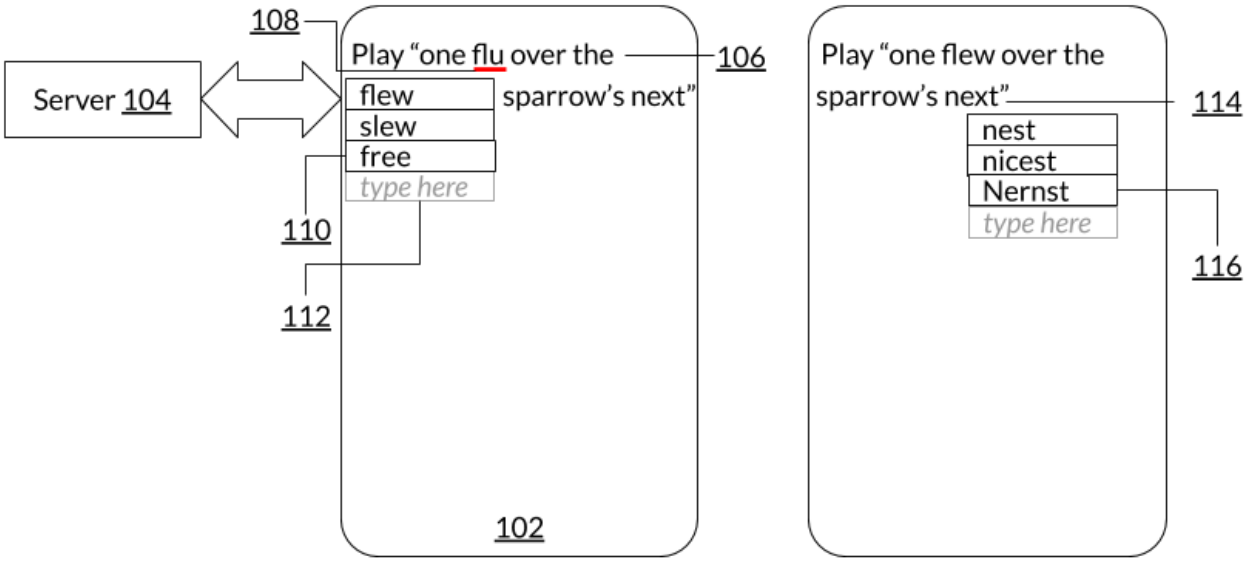

Fig. 1: Error correction in automatic speech recognizers

Fig. 1 illustrates error correction for spoken queries, per techniques of this disclosure. Speech recognition software recognizes speech received at a device (102), e.g., a smart display, a smart speaker,, a laptop, a tablet, a smartphone, or other device. Speech uttered by the user is

transcribed (106), possibly using a server (104), and displayed. Potentially misrecognized n-grams are underlined (108) or otherwise highlighted.

When the user taps on an underlined n-gram, a dropdown menu (110) provides a list of potential corrections. Entries in the dropdown menu can be provided by the speech recognition software. The user can select a suitable correction from the dropdown menu. If a suitable correction is not present in the dropdown menu, an option is provided for the user to type in the correction (112). The user can also tap and select an n-gram (114) that is not specifically identified or underlined as a potential misrecognition. If the user does so, then a dropdown menu (116) is activated with potential corrections to the selected word and a provision for the user to type in the correction. User corrections are used to further improve model performance, e.g., using speech biasing.

Identifying parts of the transcribed text that are potentially misrecognized

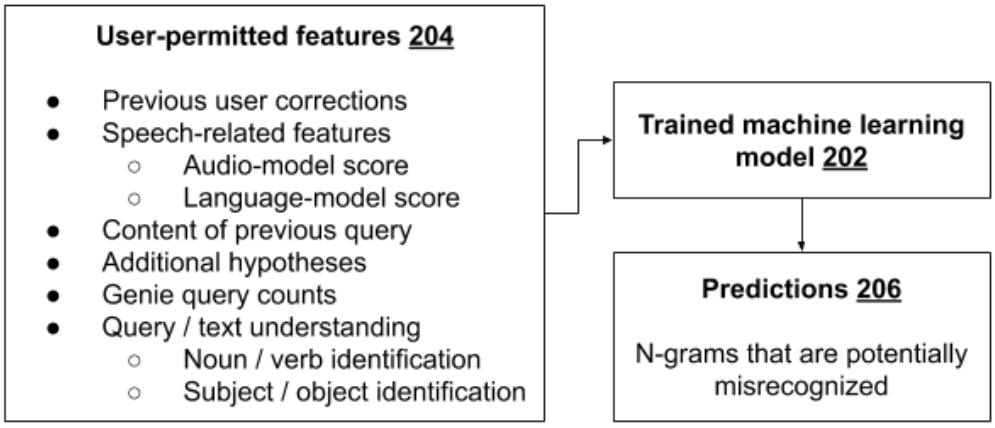

Fig. 2: Machine learning model to identify n-grams that are potentially misrecognized

Fig. 2 illustrates an example machine learning model to identify n-grams that are potentially misrecognized, per the techniques of this disclosure. A trained machine learning model (202) accepts as input user-permitted features (204) including previous user corrections;

speech-related features, e.g., audio-model scores, language-model scores, etc.; content of the previous query or utterance; additional hypotheses; genie query counts; query/text understanding, e.g., identification of nouns, verbs, subjects, and objects; etc.; and produces as output n-grams within the transcribed speech that are potentially misrecognized (208). Such n-grams, e.g., parts of the transcribed speech, are underlined or otherwise highlighted on a user interface.

Populating the dropdown menu

Correction options provided in the dropdown menu can be based upon scores produced by the ASR for various competing hypotheses for a given section of the speech. The techniques consider hypotheses arising from speech recognition and compute their differences from the underlined span or n-gram. If an overlapping delta is found, the hypothesis is added as an n-gram alternative that appears in the dropdown menu. Similarly, if the delta lies completely within the underlined span, it is added as an n-gram alternative.

Alternatively, or in addition, with user permission, correction possibilities in the dropdown menu can be based on the history of past errors that were corrected. This can be done, e.g., by aggregating speech-recognition logs into an index keyed by language/n-gram and containing other possible similar-sounding n-grams. This index can also be queried with the user-tapped n-gram to serve the best correction alternatives to 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 spoken queries, corrections made to past queries, 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 correct errors in automatic speech recognition, e.g., as performed to recognize spoken queries from a user to a virtual assistant or other application. A machine learning model detects potentially misrecognized n-grams within transcribed text which are then underlined in a user interface. A user can tap on the underlined n-gram, or another portion of the transcribed text to activate a dropdown menu that presents alternatives to the transcribed text. The alternatives can be based on speech hypothesis scores. To correct the error in transcribed text, the user picks an alternative from the dropdown menu, or, in the absence of a suitable alternative, types in the correction. With user permission, the error and corresponding correction are used as training data to improve model performance.

REFERENCES

[1] Harwath, David, Alexander Gruenstein, and Ian McGraw. "Choosing useful word alternates for automatic speech recognition correction interfaces." In *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.