January 02, 2019

Sound-based Classifier for Domain Specific Language Model Selection

Anonymous

Sound-based Classifier for Domain Specific Language Model Selection

ABSTRACT

Providing correct responses to spoken commands issued to voice-activated devices depends on such devices recognizing the commands accurately. General language models that are agnostic to functional domains have lower quality of speech recognition. An automatic speech recognition system described in this disclosure uses a sound-to-domain classifier to determine probabilities that a particular spoken phrase corresponds to particular functional domains. A language model selector picks a combination of domain-specific language models based on the probabilities. Speech recognition is performed using the selected models.

KEYWORDS

speech recognition; smart speaker; home speaker; virtual assistant; spoken command; language model; deep learning; query domain

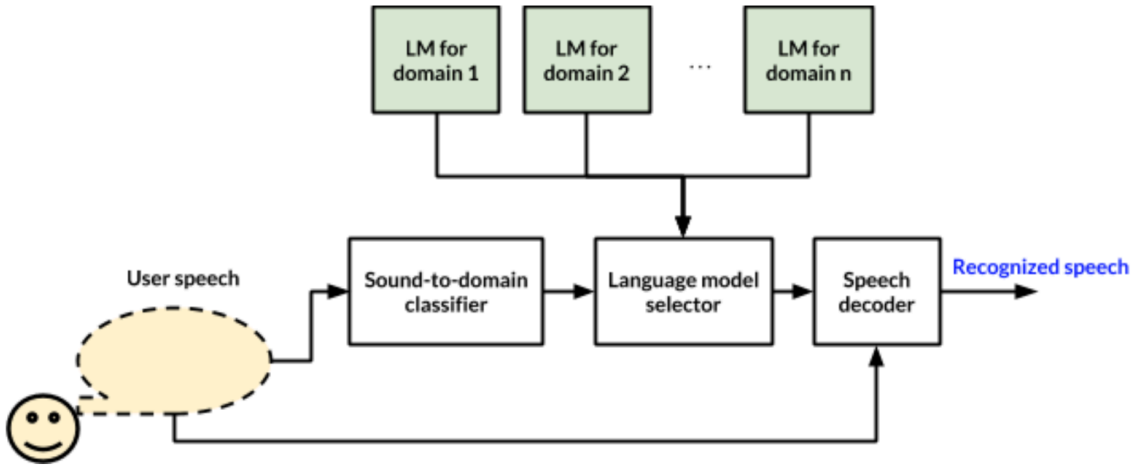
BACKGROUND

Automatic speech recognition (ASR) techniques are utilized in many applications that require computers to interpret human speech. For example, smartphones, tablets, PCs, and smart home devices such as smart speakers, video calling devices, etc. allow users to provide voice input. Most such devices provide functionality such as placing or receiving audio/video calls, music and video playback, weather information, timer and alarm functionality, device control, etc. Each functionality is part of a functional domain that has its own associated language and vocabulary.

ASR techniques that utilize a general language model (LM) and are agnostic to functional domains have many drawbacks. A general LM has a much larger search space and is therefore

often unable to provide high quality speech recognition for all the individual functional domains. A general LM requires greater computing resources. Provision of such resources may be infeasible on certain devices or in certain contexts, e.g., due to cost constraints, battery constraints, etc. Also, when functional domains overlap with each other the overall performance of the general LM can be unsatisfactory.

DESCRIPTION



This disclosure describes automatic selection of domain-specific language models based on user speech. The figure above illustrates an example automatic speech recognition system. When a user utters a word or phrase, the detected sound wave is provided to a sound-to-domain classifier.

The classifier can be implemented using any suitable techniques, e.g., a deep learning model that uses long short-term memory (LSTM) nodes. The model can be trained using suitable input, e.g., recorded audio phrases. The model can utilize different features of the audio, e.g., frequencies, phonics, and other features. Based on the features of the input user speech, the deep learning model provides as output respective probabilities of different functional domains (e.g.,

calling, music playback, weather, etc.) as being relevant to the user uttered speech. The determined probabilities are provided to the language model selector.

The ASR system includes multiple different language models, e.g., a domain-specific language model for each of the different functional domains. A language model, trained for a specific functional domain, can be applied to decode the user uttered phrase and determine a domain specific meaning.

For example, a “calling domain” language model decodes a user utterance “call John” as a command to place a call (e.g., via telephony or IP networks) to a contact named John. Different models can assign different meanings to similar user utterances. For example, user uttered phrase “Play Beyoncé” may be interpreted by a “audio playback domain” model as a command to play songs by Beyoncé, while a similar utterance “Play voicemail” may be decoded by a “messaging domain” model as a command to access the user’s voicemail.

A language model selector picks a combination of domain-specific language models based on the probabilities determined by the sound-to-domain classifier. For example, the probabilities can be used to compute respective weights for the different domains. The selected combination of models, with the assigned weights, is used to decode the user speech. The speech decoder applies the selected models and produces as its output the recognized user speech, e.g., a text version of the user uttered phrase. Use of the combination of language models narrows the search space to the specific domains of the selected models. This improves quality of the result (recognized speech) produced by the decoder.

The ASR system is scalable and supports updates to domains, e.g., addition of new domains, modifications to existing domains, deletion of domains, etc. Such updates do not negatively impact the quality of speech recognition.

CONCLUSION

An automatic speech recognition system described in this disclosure uses a sound-to-domain classifier to determine probabilities that a particular spoken phrase corresponds to particular domains. A language model selector picks a combination of domain-specific language models based on the probabilities. Speech recognition is performed using the selected models.