February 2020

Class-based Language Model for Domain-specific Speech Recognition

Anonymous

Class-based Language Model for Domain-specific Speech Recognition

ABSTRACT

Virtual assistants, as provided via devices such as smart displays or smart speakers utilize automatic speech recognition (ASR) techniques to interpret spoken commands. ASR is performed by the use of a language model that is trained using a training corpus that includes spoken variations of the commonly expected commands. However, such training does not account for words or phrases that are domain specific, e.g., names of contacts (used for calling), names of media items or artists (used for media playback), etc. Virtual assistants can therefore experience a high rate of failure to interpret commands. This disclosure describes techniques to automatically switch to a class-based language model (CLM) when specific command domains are detected in spoken queries. The CLM utilizes available user data, e.g., list of contacts, to constrain interpretation of spoken commands and can therefore achieve a high level of accuracy without the need for prior training. The use of CLM for query interpretation enables the virtual assistant to provide accurate responses.

KEYWORDS

- Automated speech recognition (ASR)
- Name recognition
- Machine learning
- Language model
- Class-based model
- Domain-specific model
- Call placement
- Virtual assistant
- Spoken query
- Smart display
- Smart speaker

BACKGROUND

Virtual assistants are popular for a variety of use cases. A device such as a smart display, smart speaker, smartphone, etc. that provides a virtual assistant enables users to provide spoken commands for a variety of tasks such as media playback, placing calls, queries for information (e.g., weather, sports scores, maps and directions), etc.

To perform the requested task, it is necessary that the spoken user command be interpreted. To this end, virtual assistants implement automated speech recognition (ASR) to obtain a transcript (sequence of words) of spoken commands. Once the query is interpreted, the virtual assistant can then perform the requested task, or ask the user to provide a clarification, e.g., if the query does not translate to a task the virtual assistant is configured to perform.

ASR involves steps such as acoustic processing, acoustic modeling, language modeling, etc. A statistical language model is a probability distribution over sequences of words. Language models provide context to distinguish between words and phrases that sound similar. A language model is typically obtained by the use of a training corpus that includes variations of spoken commands that are to be interpreted by the virtual assistant and additionally, commonly used words in the language(s) that the virtual assistant is to interpret. The resulting language model can then interpret these commands when provided by a user.

A language model is trained using such a corpus can perform poorly if the domain of the command is such that words that the model was not trained on are included in spoken commands. For example, when users invoke a virtual assistant to place outgoing calls (“calling domain”), the spoken command may be “Call X,” where X is the name of the contact to whom the call is to be placed. If the name of the contact is in the training corpus (e.g., which may include common English names such as John Smith, Jane Roberts, etc.) the query is interpreted correctly; however, many users have contacts that have last names that are uncommon (infrequent in the training corpus or not included in the training corpus), e.g., names that originate from or include words from other languages (e.g., Subramanian Kalapathy, Hoang Xiu, Joe Newth), query interpretation using a standard language model can have poor performance in interpreting the command.

This necessitates requesting the user to provide further clarifying input and provides a less than optimal calling experience. Other domains such as media playback etc. also suffer from similar gaps when portions of the query are not interpreted correctly. Further, it is impractical to train a language model with an entire corpus of domain-specific words of phrases.

DESCRIPTION

This disclosure describes techniques to overcome the problem of recognizing domain-specific words (e.g., names) by the use of a class-based language model (CLM). Per the techniques, a standard language model is utilized to interpret the user command and generate a text transcript. If the text transcript includes domain-specific keywords, e.g., at the start of the spoken input, a CLM is invoked immediately to interpret the remainder of the spoken input. The CLM is configured to interpret the remainder of the spoken input to look for a limited class of words, ignoring other possible interpretations.

For example, spoken input of the form “Call Hoang Xiu,” “Dial Joe Newth,” “Connect me to Subramanian Kalapathy,” etc. is interpreted as belonging to the calling domain based on the presence of keywords “call,” “dial,” connect me to ,” etc. Upon detection, the model is switched over to CLM that is trained to interpret the remainder of the command within a universe of names that are on the invoking user’s list of contacts. The model attempts to obtain a match for pronunciations and syllables that match the list. In addition, the likelihood of the user calling each at any given time can be calculated based on available contextual information to assign each contact a score. The contact list can be restricted to a subset of the user's contacts (e.g., removing contacts that are unlikely to be called) that is used for the CLM.

With use of the described techniques, failure rate for domain-specific commands can be reduced substantially for any combination of names, without the need for specific prior training.

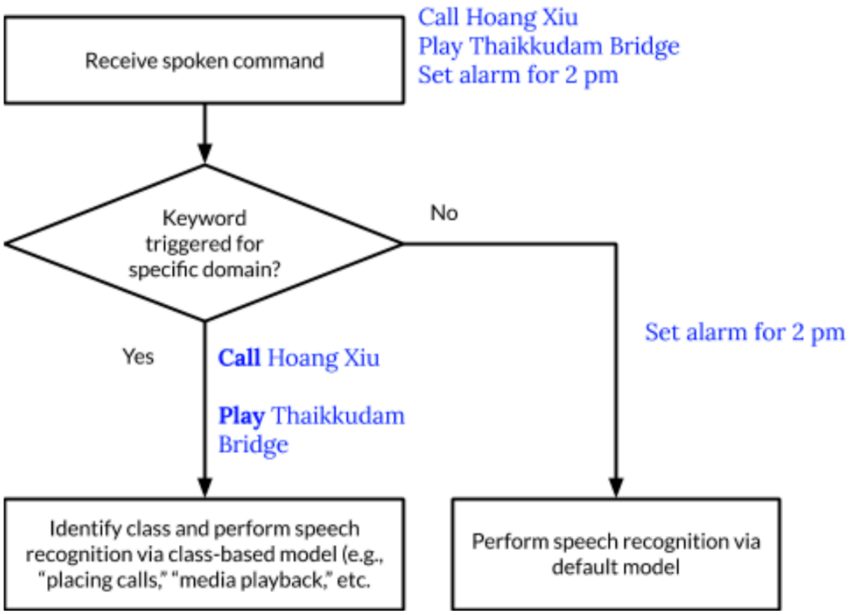
Receive spoken command
Call Hoang Xiu
Play Thaikkudam Bridge
Set alarm for 2 pm
Keyword triggered for specific domain?
No
Set alarm for 2 pm
Yes
Call Hoang Xiu
Play Thaikkudam Bridge
Identify class and perform speech recognition via class-based model (e.g., "placing calls," "media playback," etc.
Perform speech recognition via default model

Fig. 1: Use of class-based language model based on command domain

Fig. 1 illustrates a flowchart for the selective use of a CLM based on the domain of the command. A device, e.g., a smart display, is configured to receive spoken commands. When a user command is received, a standard language model is applied to parse the initial portion, of the command to determine if any domain-specific keywords for specific domains that have associated class-based models were spoken. For example, the command "Set alarm for 2 pm" is not associated with a specific domain that has a corresponding CLM, while the commands "Call Hoang Xiu" and "Play Thaikkudam Bridge" are associated with the *calling* and *media playback* domains respectively.

If the command is identified as including a keyword that triggers a domain, further speech recognition is performed using a class-based model for the specific domain; else, speech recognition is performed using a default language model. Upon completion of speech recognition, the virtual assistant can perform the requested task.

The approach described herein actively switches language models based on the initial portion of the spoken command. Thus, it can provide a high level of accuracy without requiring repeated collection of data from command failures, e.g., name recognition errors, and then training the model with such data. Available and user-permitted data can be utilized for the CLM. For example, the user’s contact list, music preferences, media playback history, etc. can be utilized to determine probabilities associated with various contacts or media items to improve interpretation of the spoken command.

CONCLUSION

This disclosure describes techniques to automatically switch to a class-based language model (CLM) when specific command domains are detected in spoken queries. The CLM utilizes available user data, e.g., list of contacts, to constrain interpretation of spoken commands and can therefore achieve a high level of accuracy without the need for prior training. The use of CLM for query interpretation enables a virtual assistant to provide accurate responses.