June 2020

Mixed Language Speech Recognition

Vladimir Vuskovic

Biraja Deo

Daisuke Ikeda

Purvi Shah

Mixed Language Speech Recognition

ABSTRACT

Users that provide spoken input mixed language are common in many geographies and application domains. Automatic speech recognition in such a context requires multiple natural language understanding (NLU) models to be run in parallel and their outputs to be combined. This disclosure describes techniques to improve the performance of such ASR models by the use of a ranking unit for language determination and assessment of whether the voice input makes sense. A response to the query is provided to the user in the language as determined by the ranking unit.

KEYWORDS

- Automatic Speech Recognition (ASR)
- Mixed language
- Multilingual
- Natural Language Understanding (NLU)
- Code-switching
- Hinglish
- Speech model
- Language detection
- 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. ASR models that are used in such contexts are trained to recognize user speech in a single language. Device users configure the device for interaction in a particular language, e.g., the primary language of the user, and the corresponding speech model is then deployed on the device for speech interaction. If the user has more than one primary language, multiple single language ASR models need to be executed in parallel to interpret the user query, since each ASR model is trained to analyze input in a particular language.

However, single language models do not work well when the user utterance that specifies the query or command includes content in more than one language in their entire command or sentence. Such phenomenon, known in linguistics as code-switching, is common in locations where users speak multiple languages and/or dialects.

For example, many users in India are bilingual in English and Hindi. The users also engage in frequent code-switching between the two languages, resulting in what is commonly known as Hinglish. An example of a mixed language query in Hinglish is "Where do I get the best *garam chai* in *Matunga*?" In this example, the frame of the query is in English, while some of the relevant elements (*garam* and *chai*), as well as the place name (*Matunga*) are in Hindi. Another example of a Hinglish mixed language query is "Mujhe *Midnight's Children* kitab kahan milegi?" In this example, the frame of the query is in Hindi, while the named entity (*Midnight's Children*) is in English. Traditional speech recognition techniques that rely on single language ASR models are insufficient to process such input where the user engages in code switching.

DESCRIPTION

This disclosure describes techniques to improve speech recognition of mixed language spoken input by leveraging language recognizers that generate multiple hypotheses as output.

The techniques improve the interpretation of spoken queries or commands where the user engages in code switching, e.g., spoken input in Hinglish or other combinations of languages.

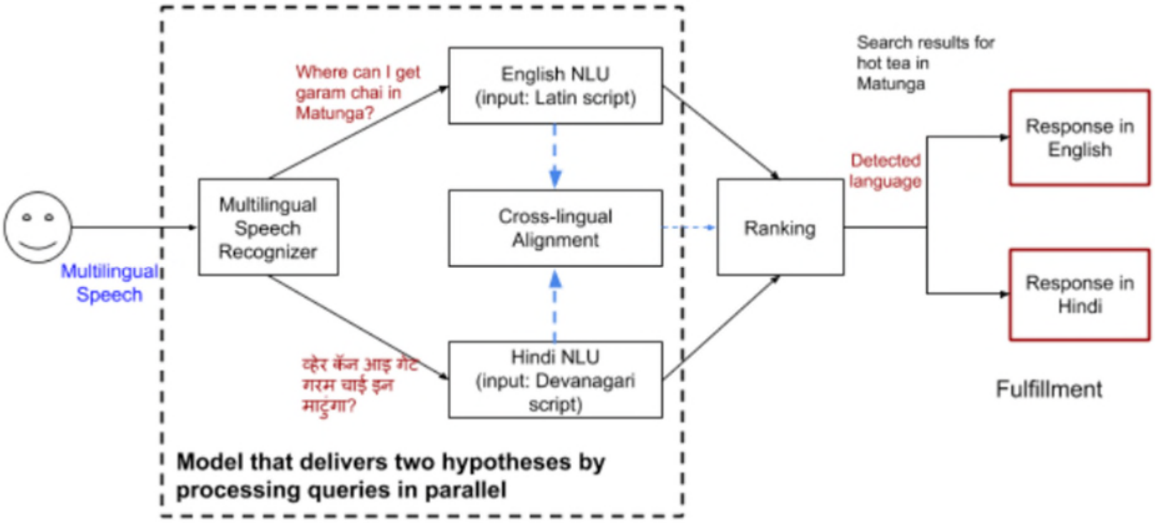
Multilingual Speech
Multilingual Speech Recognizer
Where can I get garam chai in Matunga?
English NLU (input: Latin script)
Cross-lingual Alignment
व्हेर कॅन आइ गेट गरम चाई इन माटुंगा?
Hindi NLU (input: Devanagari script)
Model that delivers two hypotheses by processing queries in parallel
Ranking
Search results for hot tea in Matunga
Detected language
Response in English
Response in Hindi
Fulfillment

Fig. 1: Mixed language speech recognition using multiple hypotheses

Per this disclosure, mixed language speech input is first processed by a multilingual speech recognizer that generates two (or more) outputs. As seen in Fig. 1, the multilingual speech recognizer provides two separate hypotheses for the two languages. While two hypotheses are shown in the example of Fig. 2, it is possible that more than two hypotheses, each corresponding to a different language may be generated. In this example, the separate hypotheses are not a transliteration of each other, but are independently determined based on interpreting the input sound in the different languages. Alternatively, the second (and additional) hypotheses can be transliterations of the first hypothesis (as shown in Fig. 1). e.g., the first output is generated as a transcription of the multilingual speech in the user's primary language, while the other outputs are transliterations of the first output in the user's secondary languages.

In Fig. 1, for the Hinglish query "Where can I get *garam chai* in *Matunga*?" the first output is in Latin script - "Where can I get garam chai in Matunga?" and the second output is in

Devanagari script - “व्हेर कॅन आइ गेट गरम चाई इन माटुंगा?” Note that while this illustrative example shows an accurate transcription, transcription of real queries may be imperfect.

Multiple natural language understanding (NLU) models are then executed in parallel, one for each of the transcriptions produced by the speech recognizer. Further, a machine translation model, e.g., implemented using neural networks or other techniques, is used to obtain translations of words and/or phrases of the spoken input. The machine translation model processes each of the transcriptions to translate the words that cannot be processed by the respective NLU model. For example, from the Latin script statement ‘Where can I get *garam chai* in *Matunga*?’, the words *garam*, *chai* and *Matunga* are translated by the machine translation model, while from the Devanagari script statement “व्हेर कॅन आइ गेट गरम चाई इन माटुंगा?”, the words व्हेर, कॅन, आइ, गेट and इन are passed to the NMT model for translation. The machine translation model provides translations, e.g., “garam=hot” and “chai=tea” respectively, for the Latin transcription. The output of each of the NLU models is updated to include the translated words received from the NMT model. In this manner, two hypotheses are generated for a Hinglish voice input as follows:

- Hypothesis 1: English Language NLU + NMT Model for cross-lingual alignment
- Hypothesis 2: Hindi Language NLU + NMT Model for cross-lingual alignment

Each of the hypotheses is assigned a confidence score, e.g., obtained based on features of the input speech (how it sounds) as well as contextual meaning of the text (whether the transcribed text makes sense). A ranking and fulfillment unit determines the suitable interpretation of the query by comparing the confidence scores of the two hypotheses and fulfills the query by providing a response to the hypothesis with the higher score. For example, for the spoken input “Where can I get garam chai in Matunga?” the response would be “Search results for hot tea in Matunga”.

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 input such as queries or commands, a user’s preferences), 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. 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 improve the performance of automatic speech recognition when the spoken input provided by a user is multilingual. The techniques utilize multiple single language ASR models in parallel to generate hypotheses for the user command. Each of the hypotheses is assigned a confidence score based on acoustic signals and contextual meaning. A ranking and fulfillment module compares the confidence scores of the two hypotheses and fulfills the query by providing a response to the hypothesis that corresponds to the higher score.

REFERENCES

1. Amazon Alexa gains support for Hindi/Hinglish conversations

2. Sung, Yun-Hsuan, Francoise Beaufays, Brian Strope, Hui Lin, and Jui-Ting Huang. "Recognizing speech in multiple languages." U.S. Patent 9,129,591, issued September 8, 2015.