FIX IT WHERE IT FAILS: PRONUNCIATION LEARNING BY MINING ERROR CORRECTIONS FROM SPEECH LOGS

Zhenzhen Kou, Daisy Stanton, Fuchun Peng, Françoise Beaufays, Trevor Strohman

Google Inc., USA

This article examines the subject of automatic speech recognition (ASR) systems and misrecognitions or incorrect pronunciations. Typically, an ASR system consists of a lexicon or word-pronunciation pairs for known word pronunciations and a grapheme-to-phoneme (G2P) engine, which generates pronunciations for unknown words. It is impossible to keep the hand-generated lexicon up to date, especially with proper names whose pronunciations are often influenced by region. The authors in this article take a unique angle in helping solve this issue, by using a language-independent approach of looking at misrecognitions and their corrections.

Decision tree classification to learn pronunciation rules, joint ngram models, maximum entropy models, active learning, and even most recently using neural networks are all areas of research that has been done on applying machine learning applications to improve G2P. Originally learning better word pronunciation through crowdsourcing was proposed for this paper; however there are a few problems to overcome, including some locally recognized but not straight forward pronunciations being unknown to those outside of the geographic region. Therefore, the authors decided to take a different route and instead learn pronunciations from recognitions correction data. Others have used acoustic and prosodic feature, examined features of the speaker’s style to detect errors, used decision-tree based methods to detect retries, and even used a co-occurrence method for detecting and correcting misrecognition. This article instead looked to use two new data mining techniques for detecting the misrecognitions: Keyboard Correction Data and Selected Alternate Data.

Keyboard Correction Data refers to query pairs (speech, keyboard) in which the user makes a voice queries and within a certain time limit (30 seconds for this article) types the same query with the keyboard. There are some challenges with this technique as the user may rephrase the query they had previously spoken when they type it in with a keyboard. The user may also choose to ask a completely unrelated question, but within the time limit. To account for these situations the correction data was classified using four feature-based criteria: word-based features, character-based features, phoneme features, and acoustic features.

Selected Alternate Data refers to a user interface that is part of the Google voice search, which allows the user to select a corrected word from a list of possible alternate results. This provides high quality corrections, as the user knows what word they were trying to communicate and presumably know how the word should be pronounced or how they want it to be pronounced. This also eliminates the need for determining bad pronunciations ahead of time since the user will identify the misrecognitions themselves.

Evaluation of the two data mining techniques for detecting misrecognitions included metrics on the word error rate (WER) on test sets containing speech queries randomly selected from traffic logs and human transcriptions. Even though many of the words in this random set would already have good pronunciations since they were from a production-level system, the authors found this data useful as a verification that their system improvements/updates had not unlearned any well-known pronunciations. In addition to the WER experiments, they also performed side-by-side (SxS) tests. Side-by-side experiments are useful because they focus on cases where pronunciation changes affect the recognition results. For the SxS experiments two ASR were built, one with the learned pronunciations and one without. Queries results for the recognition transcripts that differed were hand rated into one of four categories: nonsense – the transcript is nonsense, unusable – the transcript does not correspond to the audio, usable – the transcript contains only small errors, and exact – the transcript matches the spoken audio exactly.

Results of the experiments completed across 11 different languages showed a small reduction in WER for the Keyboard Correction Data as well as improvements in the SxS score. For the Alternate Selection Data, there was a smaller amount of data to evaluate (as less people use the interface than typing in the query themselves), so unfortunately the results showed no difference in the WER. However, there were significant improvements in the SxS evaluation going from a 0.376 to 0.432 for the English language and similar results for the other languages.

In my opinion, the authors chose a great way to learn pronunciation in using user initiated corrections. Using corrections by the user who initiated the voice query instead of guessing what the user may have intended to say helps make the results of the highest quality. Something I would be interested to see the authors look into is homophones and how they might affect results. For example, there, they’re, and there are all pronounced the same but are spelled differently and have different meanings. Unfortunately, there are a fair number of people who do not know when and/or where to use each of the different words and may use the incorrect word in their misrecognition correction. This may cause unexpected results. In the future, I would be interested to see if the data obtained from both the Keyboard Corrections and the Selected Alternates can be used to create better Selected Alternates, tightening down the corrections even more. I would also be interested to see if the algorithms used could be different for different regions. In effect, detecting either where the person currently is located or the area they are from and using a different algorithm in per region for better recognition.