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

ICASSP (2015) – 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Date of Conference: 19-24 April 2015
Page(s): 4619-4623
DOI: 10.1109/ICASSP.2015.7178846

OVERVIEW

- Automatic speech recognition (ASR) systems include a pronunciation dictionary (or lexicon) and a grapheme to phoneme (G2P) engine
  - The lexicon consists of word-pronunciation pairs written by linguist
    - Hand-generated – cannot keep up with growing vocabulary
  - G2P has limited accuracy
    - Proper names – pronunciation can be influenced by historical or foreign-origin factors
- Speech recognition task in general finds the word sequence that has the maximum posterior probability given the acoustic observations
  - Relies heavily on a lexicon
  - Uses a G2P for words not found in the lexicon

A **maximum a posteriori probability (MAP)** estimate is a mode of the posterior distribution.
The **posterior probability** of a random event or an uncertain proposition is the conditional probability that is assigned after the relevant evidence or background is taken into account. Similarly, the **posterior probability distribution** is the probability distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained from an experiment or survey.
"Posterior", in this context, means after taking into account the relevant evidence related to the particular case being examined.

EXISTING/RELATED WORK

- Research on machine learning for G2P conversion:
  - Decision tree classifier to learn pronunciation rules
  - Joint ngram model
  - Maximum entroy model
  - Active learning
  - Recurrent neural network
- Studies on detection speech recognition errors:
  - Using acoustic and prosodic features to identify corrections
  - Prosodic features to detect recognition errors
  - Examining features related to the user's speaking style to detect speech errors
  - Decision-tree based method to detect voice query retires
  - Co-occurence method for detecting and correcting misrecognition

NEW APPROACH

- Correction data focuses specifically on the areas of weaknesses of the system
  - Do not need to identify bad pronunciations ahead of time
- Language-independent
- Corrections are provided by the users who spoke them, who know how they want to pronounce the words
- Using two different types of correction data:
  - Keyboard Correction
  - Selected Alternate

KEYBOARD CORRECTION DATA

- User makes a voice query, then issues a typed query shortly after
  - Within 30 seconds
- Analysis showed only 30-40% of these pairs are true corrections
- Correction data classifier features
  - Word-based:
    - Unigram counts, number of word overlaps, and language model costs
  - Character-based:
    - Character counts, and edit distance between the recognized and typed queries
  - Phoneme:
    - Counts and edit distance between the phoneme sequences corresponding to the recognition results and typed query
  - Acoustic:
    - Forced phone alignment costs
    - Waveform-to-transcript length ratio

Table 1. Sample Keyboard Corrections (en-US)

Speech recognizer transcript	User keyboard query
Plus would Newton Kansas	Pluswood Newton Kansas
the cruise movie 3d	the croods movie 3d
what is a schematic	what is ischemic
the trucking part Arizona	the trotting park Arizona
Christmas Alex	crispus attucks

SELECTED ALTERNATE DATA

- Google voice search user interface allows users to manually select from a list of alternative recognition results
- User selection provides a high quality correction, so no extra classifier needed

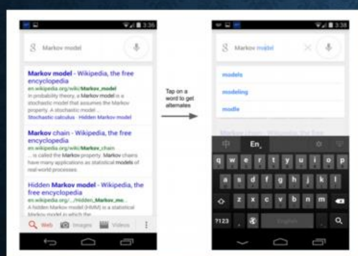


Table 2. Examples of Selected Alternates Data.

Speech recognizer transcript	User-selected alternate
my cat try to bite my Fi	my cat try to bite my thigh
which band was created in	which band was Creed in
movie tomorrow	movie Tamara
find a picture of a blue ku	find a picture of a beluga
Winston high Topeka Kansas	Quinton high Topeka Kansas
pictures of Renee swimming	pictures of Renee Fleming

TESTING

- Word error rate (WER) evaluation:
  - Anonymized speech queries randomly selected from traffic logs and human-transcribed
  - The most frequently used words already have a good pronunciation, but are still useful to ensure no learned "rogue" pronunciations
- Side-by-side (SxS) tests:
  - Two ASR engines: one with the learned pronunciations and one without
  - Both engines are fed the exact same queries from anonymized voice search logs
  - Queries with differing recognition transcripts are evaluated by human raters and marked as 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
    - Exact: the transcript matches the spoken audio exactly

WER: phone sequence for an infrequent word that matches the pronunciation of a different and more frequent word that would now be mirecognized

SxS: Acoustic model, language model, and vocabulary are the same in both engines. Only the lexicon changes.

SxS experiments have the advantage of focusing on cases where pronunciation changes do affect the recognition results. They typically show more "movement" than WER measurements on fixed test sets.

RESULTS

Keyboard Correction

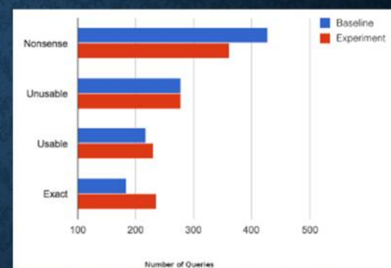
Table 3. WER comparisons for Keyboard Corrections data on American English test sets.

Data set	# Utterances	Baseline WER	Experiment WER
Search	22K	10.9	10.8
Actions	12K	21.9	21.8
Maps	18K	11.4	11.3

Table 4. SxS scores for Keyboard Corrections in three languages.

Language	Baseline score	Experiment score	p-value
English	0.376	0.432	<.001
French	0.382	0.468	<.001
German	0.416	0.489	<.001

Alternate Selection



Together

Table 5. Pronunciations learned from Keyboard Corrections and Selected Alternates

word	best G2P pronunciation	learned pronunciation
sephora	s E f O r @	s @ f O r @
tasca	t A s k @	t { s k @
estas	E s t @ z	E s t { s
verdi	v @' d i	v E r d i
newman	n u m @ n	n j u m @ n

An experiment is considered positive if its SxS score is higher than that of the corresponding baseline. Generally this means it has fewer nonsense/unusable queries and more usable and exact queries.

Keyboard Correction:
We see a small reduction in word error rate on each test set.
SxS score improvements from adding new pronunciations to the baseline ASR engine

Alternate Selection:
The amount of data flowing through the Alternate Selection pipeline is smaller than that from Keyboard Corrections. As a result, the system learns fewer pronunciations, and our experiments showed no impact on standard test set word error rates. However, SxS evaluations showed significant improvements.

Side-by-side experiments demonstrate that the pronunciations learned via our methods significantly improve the quality of a production-quality speech recognition system.

THOUGHTFUL QUESTIONS