LEARNING PRONUNCIATIONS FROM UNLABELED EVIDENCE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY

SRAVANA REDDY

CHICAGO, ILLINOIS
AUGUST 2012
Gee each owe tea eye smells fish. That’s U.

But all the fun’s in how you say a thing.

― James Joyce

― Robert Frost
# TABLE OF CONTENTS

LIST OF FIGURES ................................................................. viii
LIST OF TABLES ................................................................. ix
ACKNOWLEDGMENTS ............................................................. x
ABSTRACT ................................................................................. xii

1 INTRODUCTION .................................................................... 1
  1.1 Motivation ........................................................................ 1
    1.1.1 Why is this problem important? .................................. 1
    1.1.2 ... and why unlabeled evidence? .............................. 2
  1.2 Outline ............................................................................. 3

2 BACKGROUND ........................................................................ 5
  2.1 Grapheme-to-Phoneme Conversion ................................... 5
    2.1.1 Early Research: Local Prediction and PbA ................. 6
    2.1.2 Joint N-Gram Models ............................................. 7
    2.1.3 Discriminative Methods ......................................... 8
    2.1.4 Transliteration .................................................... 9
    2.1.5 Auxiliary Evidence in G2P .................................... 10
  2.2 Pronunciation Modeling ................................................. 10
  2.3 Inducing Phonological Categories ................................... 11
  2.4 Grapheme-to-Phoneme Conversion with Graphone FSTs ...... 11
    2.4.1 EM Learning of Graphone N-Grams ....................... 12
    2.4.2 Finite State Representation .................................. 13
    2.4.3 Experimental Results ......................................... 15
    2.4.4 Why FSTs? ..................................................... 16

3 LEARNING FROM UNTRANSCRIBED SPOKEN LANGUAGE .......... 17
  3.1 Introduction .................................................................... 17
  3.2 Previous Work .................................................................. 18
    3.2.1 Unsupervised Acoustic and Language Model Adapation . 18
    3.2.2 Out of Vocabulary Handling in ASR ....................... 19
    3.2.3 Pronunciation Learning from Transcribed Speech ....... 19
    3.2.4 Unsupervised Pronunciation Adaptation .................. 21
    3.2.5 Natural Language Processing with Non-Parallel Data .... 22
  3.3 Generative Model of Phone Sequences ............................. 22
    3.3.1 Learning Overview ............................................. 23
    3.3.2 Combining the Training Lexicon and Spoken Utterances .. 26
  3.4 Experiments .................................................................... 27
    3.4.1 Data ..................................................................... 27
    3.4.2 Evaluation .......................................................... 28
ABSTRACT

The pronunciation of a word represented in an alphabetic writing system (such as this one) is relatively transparent – but a language’s sounds change over time and vary across space, while its spellings remain relatively static, resulting in some amount of divergence between the written and spoken forms. The introduction of loanwords and proper names from other languages with different phonologies and scripts further complicates the relationship between orthography and pronunciation.

However, there are sources of information about the sound of a word in addition to spelling. This dissertation presents methods for learning pronunciations using three forms of evidence: speech, word origins, and rhymes.

Speech is the most natural source of evidence: a word’s pronunciation is usually clarified upon hearing it in a spoken utterance. In the case of proper names, knowing the linguistic or ethnic origin of the name is often instrumental in determining how it should be pronounced. Rhymes in poems or songs also provide a cue to pronunciation, particularly for inferring the sound of a word at an earlier point in history.

Using extra-orthographic evidence in a computational model necessitates access to data that provides this information in sufficiently large quantities. Collecting annotated data – speech labeled with the words that it contains, names with their ethnic origins, or poetry with rhyming patterns – can be extremely difficult and expensive. On the other hand, raw data – speech recordings, lists of names, archives of poetry – is available in plenty. The focus of this dissertation is, therefore, on using unlabeled data for pronunciation learning.

There are two facets of ‘pronunciation learning’ in this work. One is that of converting the written form of a word to its phonemic representation, known as grapheme-to-phoneme conversion. The other is a less fine-grained objective: rather than learning the exact phonemic forms, we aim to discover clusters of words with similar pronunciations.

The first contribution of this dissertation is a model to incorporate untranscribed spoken data, in addition to an existing lexicon of words and their pronunciations, into a grapheme-
to-phoneme learner. This approach is found to improve over models that are trained only on a lexicon. The second contribution is an algorithm that models the latent linguistic origins of names as part of grapheme-to-phoneme conversion. Not only does this method do better than a model without word origin awareness, it also outperforms existing approaches that first classify orthographic words by origin in a supervised manner, and then train language-specific pronunciation models. The final contribution is an unsupervised learner that discovers the rhyme schemes of stanzas of poetry, as well as rhyming relations between words and clusters of words with a common rhyming sound.