G2P Conversion of Proper Names Using Word Origin Information

Sonjia Waxmonsky and Sravana Reddy, *The University of Chicago*

Does knowing the language of origin help predict a name's pronunciation?

SH in Schoenberg (German)

S K in Schiavone (Italian)

J in Judd (English)

Y in Jung (German)

H in Juarez (Spanish)

Grapheme-to-Phoneme

Graphemes: schoenberg

Stochastic Model

Phonemes: SH OW N B ER G

Language-Aware Pronunciations

Train multiple language-specific grapheme-to-phoneme models to improve probabilistic G2P conversion

Step 1

- For words of unknown origin, train a word origin model to predict

$$\Pr(Lang \mid Word)\ \forall Lang \in L$$

Step 2

- For each *Lang*, make training dictionary containing all words where

$$\Pr(Lang \mid Word) > 0.7$$

- Train language-specific G2P model

Step 3

- *Method A:* Weight G2P output by word origin probability

$$\Pr(Pronun \mid Word) = \sum_{Lang \in L} \Pr(Lang \mid Word) \Pr(Pronun \mid Word, Lang)$$

- *Method B:* Smooth results against a language-independent model with a factor σ. In our experiments, σ=0.5 by tuning on a development set.

Data

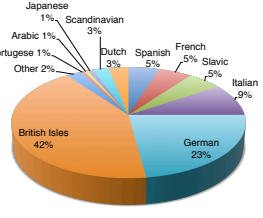
Created a corpus of common US surnames using 1990 US Census.

- Names were queried against the CMU Dictionary. Those with known pronunciations are retained, giving a set of 46k names.

- 6% of names were hand-annotated for language of origin.

- 12 most frequent languages plus "Other" class form a set of languages *L*

- Annotated data is available at people.cs.uchicago.edu/~wax/wordorigin/



A Sequential MaxEnt model is trained on hand-annotated data:

- Word represented as a character sequence

- Feature functions based on character n-grams, indicate presence of n-gram at given position in word

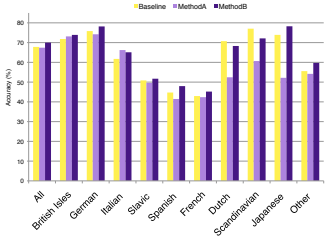
- Model is applied to remaining 46k surnames to output word origin probabilities for all names of unknown origin

Results

- The G2P algorithm used for all experiments is Sequitur[1] with 4-grams

- Data is split 80/10/10 into train/dev/test

- Baseline: fully language-independent G2P



Method A is slightly worse than the baseline because of data sparsity in language-specific G2P models

Method B is better than the baseline, because it captures language-specific pronunciations as well as Americanization patterns

Name	Baseline ✗	Method B ✓
Carcione (Italian)	K AA R S IY OW N IY	K AA R CH OW N IY
Rocha (Spanish)	R AA CH AH	R OW K AH
Wasik (Slavic)	W AA S IH K	V AA S IH K
Buescher (German)	B W EH SH ER	B Y UW SH ER
Doucet (French)	D OW S IH T	D UW S EH T

[1] Maximilian Bisani and Hermann Ney. 2008. Joint- sequence models for grapheme-to-phoneme conversion. *Speech Communication.*