G2P Conversion of Proper Names Using Word Origin Information

Sonjia Waxmonsky and **Sravana Reddy**
Department of Computer Science
The University of Chicago
Chicago, IL 60637
{wax, sravana}@cs.uchicago.edu

Abstract

Motivated by the fact that the pronunciation of a name may be influenced by its language of origin, we present methods to improve pronunciation prediction of proper names using word origin information. We train grapheme-to-phoneme (G2P) models on language-specific data sets and interpolate the outputs. We perform experiments on US surnames, a data set where word origin variation occurs naturally. Our methods can be used with any G2P algorithm that outputs posterior probabilities of phoneme sequences for a given word.

1 Introduction

Speakers can often associate proper names with their language of origin, even when the words have not been seen before. For example, many English speakers will recognize that *Makowski* and *Masiello* are Polish and Italian respectively, without prior knowledge of either name. Such recognition is important for language processing tasks since the pronunciations of out-of-vocabulary (OOV) words may depend on the language of origin. For example, as noted by Llitjós (2001), 'sch' is likely to be pronounced as /sh/ for German-origin names (*Schoenenberg*) and /sk/ for Italian-origin words (*Schiavone*).

In this work, we apply word origin recognition to grapheme-to-phoneme (G2P) conversion, the task of predicting the phonemic representation of a word given its written form. We specifically study G2P conversion for personal surnames, a domain where OOVs are common and expected.

Our goal is to show how word origin information can be used to train language-specific G2P models, and how output from these models can be combined to improve prediction of the best pronunciation of a name. We deal with data sparsity in rare language classes by re-weighting the output of the language-specific and language-independent models.

2 Previous Work

Llitjós (2001) applies word origin information to pronunciation modeling for speech synthesis. Here, a CART decision tree system is presented for G2P conversion that maps letters to phonemes using local context. Experiments use a data set of US surnames that naturally draws from a diverse set of origin languages, and show that the inclusion of word origin features in the model improves pronunciation accuracy. We use similar data, as described in §4.1.

Some works on lexical modeling for speech recognition also make use of word origin. Here, the focus is on expanding the vocabulary of an ASR system rather than choosing a single best pronunciation. Maison et al. (2003) train language-specific G2P models for eight languages and output pronunciations to augment a baseline lexicon. This augmented lexicon outperforms a handcrafted lexicon in ASR experiments; error reduction is highest for foreign names spoken by native speakers of the origin language. Cremelie and ten Bosch (2001) carry out a similar lexicon augmentation, and make use of *penalty weighting*, with different penalties for pronunciations generated by the language-specific and language-independent G2P models.

The problem of machine transliteration is closely related to grapheme-to-phoneme conversion. Many

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 367–371,
Montréal, Canada, June 3-8, 2012.

transliteration systems (Khapra and Bhattacharyya, 2009; Bose and Sarkar, 2009; Bhargava and Kondrak, 2010) use word origin information. The method described by Hagiwara and Sekine (2011) is similar to our work, except that (a) we use a data set where multiple languages of origin occur naturally, rather than creating language-specific lists and merging them into a single set, and (b) we consider methods of smoothing against a language-independent model to overcome the problems of data sparsity and errors in word origin recognition.

3 Language-Aware G2P

Our methods are designed to be used with any statistical G2P system that produces the posterior probability $\Pr(\bar{\phi}|\bar{g})$ of a phoneme sequence $\bar{\phi}$ for a word (grapheme sequence) $\bar{g}$ (or a score that can be normalized to give a probability). The most likely pronunciation of a word is taken to be $\arg\max_{\bar{\phi}} \Pr(\bar{\phi}|\bar{g})$.

Our baseline is a single G2P model that is trained on all available training data. We train additional models on language-specific training subsets and incorporate the output of these models to re-estimate $\Pr(\bar{\phi}|\bar{g})$, which involves the following steps:

1. Train a supervised word origin classifier to predict $\Pr(l|w)$ for all $l \in L$, the set of languages in our hand-labeled word origin training set.
2. Train G2P models for each $l \in L$. Each model m_l is trained on words with $\Pr(l|w)$ greater than some threshold α. Here, we use $\alpha = 0.7$.
3. For each word w in the test set, generate candidate transcriptions from model m_l for each language with nonzero $\Pr(l|w)$. Re-estimate $\Pr(\bar{\phi}|\bar{g})$ by interpolating the outputs of the language-specific models. We may also use the output of the language-independent model.

We elaborate on our approaches to Steps 1 and 3.

3.1 Step 1: Word origin modeling

We apply a sequential conditional model to predict $\Pr(l|w)$, the probability of a language class given the word. A similar Maximum Entropy model is presented by Chen and Maison (2003), where features are the presence or absence of a given character n-gram in w. In our approach, feature functions are defined at character positions rather than over the entire word. Specifically, for word w_j composed of character sequence $c_1 \ldots c_m$ of length m (including start and end symbols), binary features test for the presence or absence of an n-gram *context* at each position m. A context is the presence of a character n-gram starting or ending at position m. Model features are represented as:

$$f_i(w, m, l_k) = \begin{cases} 1, & \text{if } lang(w) = l_k \text{ and context} \\ & \quad i \text{ is present at position } m \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Then, for $w_j = c_i \ldots c_m$:

$$\Pr(l_k|w_j) = \frac{\exp \sum_m \sum_i \lambda_i f_i(c_m, l_k)}{\mathbf{Z}} \tag{2}$$

where $\mathbf{Z} = \sum_j \exp \sum_m \sum_i \lambda_i f_i(c_m, l_k)$ is a normalization factor. In practice, we can implement this model as a CRF, where a language label is applied at each character position rather than for the word.

While all the language labels in a sequence need not be the same, we find only a handful of words where a transition occurs from one language label to another within a word. For these cases, we take the label of the last character in the word as the language of origin. Experiments comparing this sequential Maximum Entropy method with other word origin classifiers are described by Waxmonsky (2011).

3.2 Step 3: Re-weighting of G2P output

We test two methods of re-weighting $\Pr(\bar{\phi}|\bar{g})$ using the word origin estimation and the output of language-specific G2P models.

Method A uses only language-specific models:

$$\tilde{\Pr}(\bar{\phi}|\bar{g}) = \sum_{l \in L} \Pr(\bar{\phi}|\bar{g}, l) \Pr(l|g) \tag{3}$$

where $\Pr(\bar{\phi}|\bar{g}, l)$ is estimated by model m_l.

Method B With the previous method, names from infrequent classes suffer from data sparsity. We therefore smooth with the output P_I of the baseline language-independent model.

$$\tilde{\Pr}(\bar{\phi}|\bar{g}) = \sigma \Pr_I(\bar{\phi}|\bar{g}) + (1-\sigma) \sum_{l \in L} \Pr(\bar{\phi}|\bar{g}, l) \Pr(l|g) \tag{4}$$

The factor σ is tuned on a development set.

Language Class	Train Count	Test Count	Base -line	(A)	(B)
British	16.1k	2111	71.8	73.1	**73.9**
German	8360	1109	75.8	74.2	**78.2**
Italian	3358	447	61.7	**66.2**	65.1
Slavic	1658	232	50.9	49.6	**51.7**
Spanish	1460	246	44.7	41.5	**48.0**
French	1143	177	42.9	42.4	**45.2**
Dutch	468	82	**70.7**	52.4	68.3
Scandin.	393	61	**77.1**	60.7	72.1
Japanese	116	23	73.9	52.2	**78.3**
Arabic	68	18	33.3	11.1	**38.9**
Portug.	34	4	25.0	25.0	**50.0**
Hungarian	28	3	**100.0**	66.7	**100.0**
Other	431	72	55.6	54.2	**59.7**
All			67.8	67.4	**70.0**

Table 1: G2P word accuracy for various weighting methods using a character-based word origin model.

4 Experiments

4.1 Data

We assemble a data set of surnames that occur frequently in the United States. Since surnames are often "Americanized" in their written and phonemic forms, our goal is to model how a name is most likely to be pronounced in standard US English rather than in its language of origin.

We consider the 50,000 most frequent surnames in the 1990 census[1], and extract those entries that also appear in the CMU Pronouncing Dictionary[2], giving us a set of 45,841 surnames with their phoneme representations transcribed in the Arpabet symbol set. We divide this data 80/10/10 into train, test, and development sets.

To build a word origin classification training set, we randomly select 3,000 surnames from the same census lists, and label by hand the *most likely* language of origin of each name when it occurs in the US. Labeling was done primarily using the *Dictionary of American Family Names* (Hanks, 2003) and Ellis Island immigration records.[3] We find that, in many cases, a surname cannot be attributed to a single language but can be assigned to a set of languages related by geography and language family. For example, we discovered several surnames that could be ambiguously labeled as English, Scottish, or Irish in origin. For languages that are frequently confusable, we create a single language group to be used as a class label. Here, we use groups for British Isles, Slavic, and Scandinavian languages. Names of undetermined origin are removed, leaving a final training set of 2,795 labeled surnames and 33 different language classes. We have made this annotated word origin data publicly available for future research.[4]

In these experiments, we use surnames from the 12 language classes that contain at least 10 hand-labeled words, and merge the remaining languages into an "Other" class. Table 1 shows the final language classes used. Unlike the training sets, we do not remove names with ambiguous or unknown origin from the test set, so our G2P system is also evaluated on the ambiguous names.

4.2 Results

The Sequitur G2P algorithm (Bisani and Ney, 2008) is used for all our experiments.

We use the CMU Dictionary as the gold standard, with the assumption that it contains the standard pronunciations in US English. While surnames may have multiple valid pronunciations, we make the simplifying assumption that a name has one best pronunciation. Evaluation is done on the test set of 4,585 names from the CMU Dictionary.

Table 1 shows G2P accuracy for the baseline system and Methods A and B. Test data is partitioned by the most likely language of origin.

We see that Method A, which uses only language-specific G2P models, has lower overall accuracy than the baseline. We attribute this to data sparsity introduced by dividing the training set by language. With the exception of British and German, language-specific training set sizes are less than 10% the size of the baseline training set of 37k names. Another cause of the lowered performance is likely due to errors made by our word origin model.

Examining results for individual language classes for Method A, we see that Italian and British are

[1] http://www.census.gov/genealogy/names/

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] http://www.ellisisland.org

[4] The data may be downloaded from http://people.cs.uchicago.edu/~wax/wordorigin/.

Language	Surname	Baseline	Method B
Italian	Carcione	K AA R S IY OW N IY	K AA R CH OW N IY
	Cuttino	K AH T IY N OW	K UW T IY N OW
	Lubrano	L AH B R AA N OW	L UW B R AA N OW
	Pesola	P EH S AH L AH	P EH S OW L AH
Slavic	Kotula	K OW T UW L AH	K AH T UW L AH
	Jaworowski	JH AH W ER AO F S K IY	Y AH W ER AO F S K IY
	Lisak	L IY S AH K	L IH S AH K
	Wasik	W AA S IH K	V AA S IH K
Spanish	Bencivenga	B EH N S IH V IH N G AH	B EH N CH IY V EH NG G AH
	Vivona	V IH V OW N AH	V IY V OW N AH
	Zavadil	Z AA V AA D AH L	Z AA V AA D IY L

Table 2: Sample G2P output from the Baseline (language-independent) and Method B systems. Language labels shown here are the $\arg\max_l P(l|w)$ using the character-based word origin model. Phoneme symbols are from an Arpabet-based alphabet, as used in the CMU Pronouncing Dictionary.

the only language classes where accuracy improves. For Italian, we attribute this to two factors: high divergence in pronunciation from US English, and the availability of enough training data to build a successful language-specific model. In the case of British, a language-specific model removes foreign words but leaves enough training data to model the language sufficiently.

Method B shows accuracy gains of 2.2%, with gains for almost all language classes except Dutch and Scandinavian. This is probably because names in these two classes have almost standard US English pronunciations, and are already well-modeled by a language-independent model.

We next look at some sample outputs from our G2P systems. Table 2 shows names where Method B generated the gold standard pronunciation and the baseline system did not. For the Italian and Spanish sets, we see that the letter-to-phoneme mappings produced by Method B are indicative of the language of origin: (c → /CH/) in *Carcione*, (u → /UW/) in *Cuttino*, (o → /OW/) in *Pesola*, and (i → /IY/) in *Zavadil* and *Vivona*. Interestingly, the name *Bencivenga* is categorized as Spanish but appears with the letter-to-phoneme mapping (c → /CH/), which corresponds to Italian as the language of origin. We found other examples of the (c → /CH/) mappings, indicating that Italian-origin names have been folded into Spanish data. This is not surprising since Spanish and Italian names have high confusion with each other. Effectively, our word origin model produced a noisy Spanish G2P training set, but the re-weighted G2P system is robust to these errors.

We see examples in the Slavic set where the gold standard dictionary pronunciation is partially but not completely Americanized. In *Jaworowski*, we have the mappings (j → /Y/) and (w → /F/), both of which are derived from the original Polish pronunciation. But for the same name, we also have (w → /W/) rather than (w → /V/), although the latter is truer to the original Polish. This illustrates one of the goals of our project, which is is to capture these patterns of Americanization as they occur in the data.

5 Conclusion

We apply word origin modeling to grapheme-to-phoneme conversion, interpolating between language-independent and language-specific probabilistic grapheme-to-phoneme models. We find that our system outperforms the baseline in predicting Americanized surname pronunciations and captures several letter-to-phoneme features that are specific to the language of origin.

Our method operates as a wrapper around G2P output without modifying the underlying algorithm, and therefore can be applied to any state-of-the-art G2P system that outputs posterior probabilities of phoneme sequences for a word.

Future work will consider unsupervised or semi-supervised approaches to word origin recognition for this task, and methods to tune the smoothing weights σ at the language rather than the global level.

References

Aditya Bhargava and Grzegorz Kondrak. 2010. Language identification of names with SVMs. In *Proceedings of NAACL*.

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

Dipankar Bose and Sudeshna Sarkar. 2009. Learning multi character alignment rules and classification of training data for transliteration. In *Proceedings of the ACL Named Entities Workshop*.

Stanley F. Chen and Benoît Maison. 2003. Using place name data to train language identification models. In *Proceedings of Eurospeech*.

Nick Cremelie and Louis ten Bosch. 2001. Improving the recognition of foreign names and non-native speech by combining multiple grapheme-to-phoneme converters. In *Proceedings of ITRW on Adaptation Methods for Speech Recognition*.

Masato Hagiwara and Satoshi Sekine. 2011. Latent class transliteration based on source language origin. In *Proceedings of ACL*.

Patrick Hanks. 2003. Dictionary of American family names. *New York : Oxford University Press*.

Mitesh M. Khapra and Pushpak Bhattacharyya. 2009. Improving transliteration accuracy using word-origin detection and lexicon lookup. In *Proceedings of the ACL Named Entities Workshop*.

Ariadna Font Llitjós. 2001. Improving pronunciation accuracy of proper names with language origin classes. Master's thesis, Carnegie Mellon University.

Benoît Maison, Stanley F. Chen, and Paul S. Cohen. 2003. Pronunciation modeling for names of foreign origin. In *Proceedings of ASRU*.

Sonjia Waxmonsky. 2011. *Natural language processing for named entities with word-internal information.* Ph.D. thesis, University of Chicago.