Language Identification and Language Specific Letter-to-Sound Rules

Stephen Lewis, Katie McGrath, Jeffrey Reuppel

University of Colorado at Boulder

This paper describes a system that improves automatic ARPABET transcription by addressing performance issues resulting from Arabic and Russian transliteration in English text. Our system is called EAR (English, Arabic, Russian). The EAR system has two components: 1. An n-gram language identifier module which classifies an incoming unknown word as Arabic, Russian, or English, 2. Language specific letter to sound rules which output a pronunciation for a word based on its classification. Our results show overall system error reduction rates at upwards of 45% as compared to a system trained only on English.

1. Introduction

The sparsity of transcribed conversational English makes assembling a corpus for speech recognition training a challenging task. One alternative resource for dealing with sparsity is to mine the World Wide Web for transcribed conversations. Utilizing this resource though poses a number of new problems from text normalization to producing optimized output for letter to speech.

This transcribed English text, especially news text, contains significant a number of transliterated foreign proper nouns. Without normalization the output of a text-to-speech system will often generate inappropriate pronunciations because the letter-to-sound rules have been trained on English spelling standards, not on foreign transliteration standards. Applying English letter-to-sound rules to Arabic transliteration, for instance, can result in the following mispronunciation.

Example 1.

Hibaaq → HH AY B AE KD[1]

To address the issue of poorly produced pronunciation data due to the presence of non-English words, we have implemented a two-tiered language classifier. The first tier of this classifier serves to identify non-English words in the conversational data scraped from the web. The second tier processes the results of this classification through specific letter-to-sound rules that have been trained for each non-English language in question. More specifically, our research has focused on solving these problems for Arabic and Russian words in our web scraped corpus. Words from both of these languages had

[1] Wrong letter to sound output of Arabic Word expanded into ARBABET symbols using Decision trees trained on CMU's English lexicon. For details about ARPABET symbols, see (JURAFSKY & MARTIN 2000, p 94-95). ARPABET [HH AY B AE KD] is roughly equivalent to IPA [□□□□ □].

Colorado Research in Linguistics. June 2004. Volume 17, Issue 1. Boulder: University of Colorado.

already been identified as both common and problematic to the production of training material.

2. Past/Related Work

Off the shelf technology which can perform language identification on text in general is available. These same tools can also be used to produce letter to speech pronunciation output for non-English words. However none of these systems are appropriate for the task that we wish to accomplish. Most existing language classification systems use the simple trick of determining the character encoding of a document to perform language identification. A few systems use n-gram statistics at the word level to perform this task.

Both of these techniques can be used to identify non-English words successfully. However this identification works within the context of the entirety of the text in question being composed in the non-English language. This type of system is not optimized to deal with identifying the origin of non-English words within an otherwise English language text.

The same is often true of letter-to-sound systems. They are built to pronounce non-English words correctly within the context of the language of origin. However the pronunciations are representative of native speaker pronunciation and inappropriate to producing the letter to sound output indicative of a monolingual English speaker attempting to pronounce an unknown word of foreign origin.

3. Language Classification

3.1. Training Data

Since Arabic and Russian proper nouns had previously been identified as the primary cause for pronunciation errors our first task was to acquire training data by which to identify these words. Using the World Wide Web as our resource we constructed a database of transliterated proper nouns in Arabic and Russian names. The Arabic names were obtained on the web in numerous locations to create a list of 3143 unique Arabic names. The Russian names were all obtained from a single web site that provided 20577 unique Russian names (GOLDSCHMIDT 1996). In order to create a standard of comparison for English words, we also collected a list of the 10,000 most common words longer than 4 letters from the Brown Corpus. We collected English words from the Brown Corpus with the belief that such an assemblage would better represent the body of unknown English words. While new foreign words are generally proper nouns, "unknown" English words tend to be morphological variants of words in the lexicon. In addition, English first names are not at all representative of the greater language, as evidenced by the 5 most common 4-grams from a list of English names and the Brown Corpus (Table 1).

DOI: https://doi.org/10.25810/60mf-fn94

Names	Brown Corpus
<s>MAR	ING</s>
ANA</s>	TION
NNA</s>	ION</s>
INA</s>	ATIO
ANNA	TED</s>

Table 1. The 5 most common 4-grams from a list of English names and from the Brown Corpus

3.2. Classification Algorithm

We implemented an n-gram classifier to handle language type identification. Each training word was segmented into individual letters. Individual 4-grams were constructed using each four-letter set. In addition much like sentence boundaries are marked, letters which begin and end words were marked with <s> and <\s>. Each individual 4-gram was then assigned a specific probability based on frequency. The word to be classified was also segmented into 4-grams and then labeled for language using the following equation (Equation 1).

Equation 1:

$$C = \text{argmax}_c P(c|x_1,\ldots,x_n) =$$

$$\text{argmax}_c \; P(c) \prod_{i=1} P(x_i|c)$$

C – final language classification

c – individual language classification

x – 4-gram

n – number of 4-grams in the word being classified

Prior probabilities for each language were generated in proportion to the content of the CNN corpus. This set of news transcriptions was then compared against the CMU lexicon. This comparison returned a list of 1001 "unknown" words. Each unknown word was labeled by a team of linguistics graduate students as being Arabic, Russian, or other. Each word was then classified according to the majority label given it and the percentages of each language classification determined the priors. The priors were set as in Table 2:

English: 0.805
Arabic: 0.156
Russian: 0.039

Table 2. Prior probabilities for labeling words English, Arabic, or Russian

As with any n-gram classifier, we needed a way to calculate probabilities for new 4-grams that never appeared in the training data. To account for these unseen 4-grams, we used a modified add-one smoothing method Using this method, the probability for each unseen 4-gram was assigned to be the same as that of the 4-grams with the lowest overall probability. In the same 1001 unknown words mentioned above, unseen 4-grams accounted for only 0.126% of all 4-grams. As smoothing is invoked so infrequently, more sophisticated forms of smoothing or back-off do not seem to be fertile avenues to travel down for system improvement.

4. Letter to Sound Rules

4.1. Training Data

Our letter to sound output systems for Arabic and Russian are built upon the integration of two separate resources. First our lexicon is structured in the same way as the CMU pronunciation dictionary reformatted to sphinx format (CMU Sphinx). Second we have used the SONIC (PELLOM, 2003) decision tree software to train our the Letter-to-Sound rules on producing the correct output consisting of a word and it's ARPABET transcription.

Given the absence of a proper corpus of Arabic and Russian words transcribed in this manner, we once again scraped the web for data. After collecting transliterated words from various resources theses words were hand-transcribed into ARPABET phonemic output by a team of graduate linguists. In all we built two corpora of transliterated words, 844 Russian words and 582 Arabic words. This corpus is small, but since the words were hand-transcribed, its data is at least reliable. Both transcription time and the previous shortage of appropriately transliterated words contributed to the decision to use a small but reliable data set.

4.2. Training Algorithm

For our decision tree training algorithm we used borrowed technology from the SONIC system. The SONIC system allows us to produce Letter-to-Sound output for words which we have never seen before and which do not exist in our language specific training corpora.

DOI: https://doi.org/10.25810/60mf-fn94

The algorithm works by extracting feature vectors from the input data consisting of the center letter plus 3 letters of context. Each output phoneme is selected using a greatest reduction of entropy measure.

Data preparation for using the SONIC system requires input and produced output as shown in example 2.

Example 2:

Input

GOLOVA G OW L AX V AA

Output

K AA R IY K UW (CARICU)

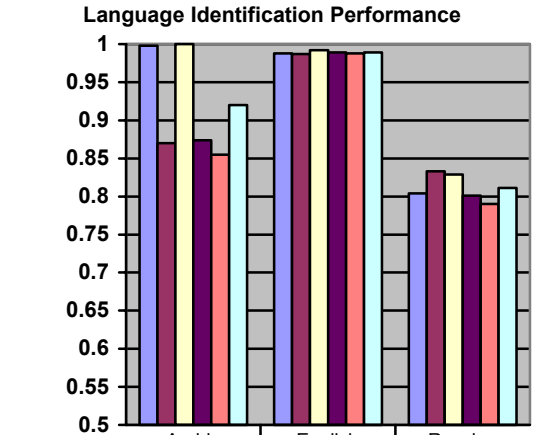
5. Results/Analysis

Analysis of the system was done in 3 distinct phases – analysis of the classifier, analysis of the letter-to-sound rules, and analysis of the complete system

5.1. Phase 1: Analysis of the classifier

The language identification classifier training data was split to create test data from 20% of each of the three language word lists. The classifier was then trained using the remaining 80% of each list. This process was repeated 4 times with the training data split differently each time, providing 5 different overlapping training sets and test sets. Each was analyzed for precision by simply counting the number of times each classifier correctly labeled the words in the test sets for each language. The classifier's precision on the Arabic test data sets described above ranged from .86 to 1.0 with a mean of .92.

On the Russian data precision ranged from .79 to .83 with a mean of .81. The classifier achieved its highest accuracy and greatest consistency on the English data with a precision ranging from .986 to .992 and a mean of .988.

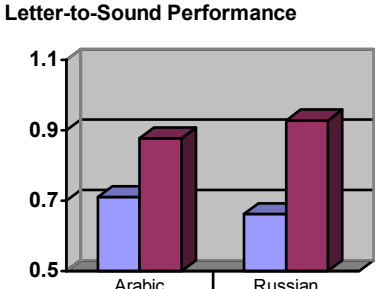


	Arabic	English	Russian
classifier 1	0.998	0.988	0.804
classifier 2	0.87	0.987	0.833
classifier 3	1	0.992	0.829
classifier 4	0.874	0.989	0.801
classifier 5	0.855	0.988	0.79
mean	0.92	0.989	0.811

Figure 1. Performance of the classifier as trained and tested on 5 overlapping data sets (classifiers 1-5)

5.2. Phase 2: Analysis of the letter-to-sound rules

Of the hand transcribed Arabic and Russian transliterated words, 10% were reserved for testing the language specific letter-to-sound rule decision trees. Performance of the language specific decision trees was compared against the performance of a decision tree trained on the CMU lexicon as a baseline. Accuracy was determined by counting the phones that matched between the hand transcription and the decision tree transcription. The Russian decision tree performed at a precision of .93 (error rate .07) on the Russian test set. The baseline decision tree performed at a precision of .66 (error rate .33) on the



	Arabic	Russian
Baseline (CMU)	0.711	0.662
Language Specific	0.877	0.928

Figure 2. Performance of language specific letter to sound rules on language specific data as compared to baseline

same test set, showing a 79 percent reduction in error rate by using the Russian decision tree. The Arabic decision tree performed at a precision of .88 (error rate .12) on the Arabic test set. The baseline decision tree performed at a precision of .71 (error rate .29) on the same test set, showing a 57 percent reduction in error rate by using the Arabic decision tree.

5.3. Phase 3: Analysis of the complete system

To test the complete system, we isolated 50 words from each hand transcribed test set. We then isolated 50 English words from the CMU lexicon, and combined the three sets, making 150 transcribed words, 50 in each language. These words were classified using the language identification classifier and then transcribed using appropriate decision trees based on the classifier labels. Counts were then made of the phones in the system's transcription that matched the hand transcription and lexicon transcription. On this data, the system achieved a precision of .89. The decision tree trained on the CMU lexicon performed with a precision of .80 on the same data, showing a .46 reduction in error rate by using the new system.

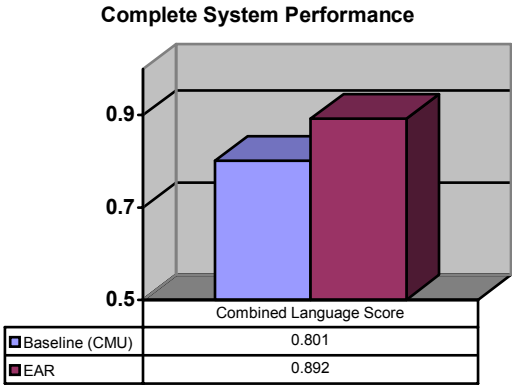


Figure 3. Performance of the language identification classifier and the language specific letter-to-sound rules combined on 50 English, Russian, and Arabic words

6. Future Work and Conclusions

Our results are quite promising. We have shown that we can significantly improve automatic transcription by integrating an n-gram language classifier and language specific transcription decision trees.

These promising results warrant further experimentation. There is a wide range of machine learning techniques which we would like to implement in the future as an alternative to using simple n-grams for the first tier of our classifier. We would like to thank fellow researcher Dan Cer for his suggestion that an unsupervised machine learning approach could be applied to this classification task. Preliminary investigations imply that unsupervised ML is promising.

It is likely that we will see further improvement in the performance of our present day classifier by improving the quality of our training data. Suggestions for such improvements have included the addition of non-name based Russian and Arabic data, comparable to the English Brown corpus training data being used. It may also be possible to improve performance by utilizing a filtered Russian-name data set, pared down from the large and potentially noisy set we are currently using.

The algorithms used in generating the letter-to-sound rules are those commonly accepted as standards. However, we expect that augmenting the size of our transcribed non-English data would increase the quality of output. This will be one of the first areas we will seek to retune as it requires only getting access to more transliterations, transcribing them into phonemes, and incorporating this new data into our training set.

Finally, our system could easily be extended to handle other non-English word sets via the implementation of additional language specific LTS classifiers, and by adding additional classification groups for categorizing among a different set of languages.

The functioning of our system requires that there exist a degree of statistical distinction between languages in order to effectively differentiate between them. Future work could easily address the theoretical question of assessing the limits on the number of languages it can handle based upon their phonemic similarity and difference. Additionally this would serve to test whether the system is robust enough to distinguish between more closely related languages.

References

COKER, K. CHURCH AND M. LIBERMAN. 1990. 'Morphology and Rhyming: Two Powerful Alternatives to Letter-to-Sound Rules for Speech Synthesis.' European Speech Communication Association, Conference on Speech Synthesis.

JURAFSKY, DANIEL AND JIM MARTIN. 2000. *Speech and Language Processing*. Prentice Hall.

PELLOM, BRYAN AND KADRI HACIOGLU. 2003. SONIC: The University of Colorado Continuous Speech Recognizer, Technical Report TR-CSLR-2001-01.

REYNAR, JEFFREY C. AND ADWAIT RATNAPARKHI. 1997. 'A Maximum Entropy Approach to Identifying Sentence Boundaries.' In Proceedings of the Fifth Conference on Applied Natural Language Processing.

RUSSELL, STUART AND PETER NORVIG. 2002. *Artificial Intelligence: A Modern Approach*. Prentice Hall.

URLs

GOLDSCHMIDT, PAUL. 1996. A Dictionary of Period Russian Names. http://www.sca.org/heraldry/paul/. Retrieved December 2003.

CMU SPHINX. http://fife.speech.cs.cmu.edu/sphinx/. Retrieved December 2003.

DOI: https://doi.org/10.25810/60mf-fn94