Does the Phonology of L1 Show Up in L2 Texts?

Garrett Nicolai and **Grzegorz Kondrak**
Department of Computing Science
University of Alberta
{nicolai,gkondrak}@ualberta.ca

Abstract

The relative frequencies of character bigrams appear to contain much information for predicting the first language (L1) of the writer of a text in another language (L2). Tsur and Rappoport (2007) interpret this fact as evidence that word choice is dictated by the phonology of L1. In order to test their hypothesis, we design an algorithm to identify the most discriminative words and the corresponding character bigrams, and perform two experiments to quantify their impact on the L1 identification task. The results strongly suggest an alternative explanation of the effectiveness of character bigrams in identifying the native language of a writer.

1 Introduction

The task of Native Language Identification (NLI) is to determine the first language of the writer of a text in another language. In a ground-breaking paper, Koppel et al. (2005) propose a set of features for this task: function words, character n-grams, rare part-of-speech bigrams, and various types of errors. They report 80% accuracy in classifying a set of English texts into five L1 languages using a multi-class linear SVM.

The First Shared Task on Native Language Identification (Tetreault et al., 2013) attracted submissions from 29 teams. The accuracy on a set of English texts representing eleven L1 languages ranged from 31% to 83%. Many types of features were employed, including word length, sentence length, paragraph length, document length, sentence complexity, punctuation and capitalization, cognates, dependency parses, topic models, word suffixes, collocations, function word n-grams, skip-grams, word networks, Tree Substitution Grammars, string kernels, cohesion, and passive constructions (Abu-Jbara et al., 2013; Li, 2013; Brooke and Hirst, 2013; Cimino et al., 2013; Daudaravicius, 2013; Goutte et al., 2013; Henderson et al., 2013; Hladka et al., 2013; Bykh et al., 2013; Lahiri and Mihalcea, 2013; Lynum, 2013; Malmasi et al., 2013; Mizumoto et al., 2013; Nicolai et al., 2013; Popescu and Ionescu, 2013; Swanson, 2013; Tsvetkov et al., 2013). In particular, word n-gram features appear to be particularly effective, as they were used by the most competitive teams, including the one that achieved the highest overall accuracy (Jarvis et al., 2013). Furthermore, the most discriminative word n-grams often contained the name of the native language, or countries where it is commonly spoken (Gebre et al., 2013; Malmasi et al., 2013; Nicolai et al., 2013). We refer to such words as *toponymic terms*.

There is no doubt that the toponymic terms are useful for increasing the NLI accuracy; however, from the psycho-linguistic perspective, we are more interested in what characteristics of L1 show up in L2 texts. Clearly, L1 affects the L2 writing in general, and the choice of words in particular, but what is the role played by the phonology? Tsur and Rappoport (2007) observe that limiting the set of features to the relative frequency of the 200 most frequent character bigrams yields a respectable 66% accuracy on a 5-language classification task. The authors propose the following hypothesis to explain this finding: "*the choice of words* [emphasis added] people make when writing in a second language is strongly influenced by the phonology of their native language". As the orthography of alphabetic languages is at least partially representative of the underlying phonology, character bigrams may capture these phonological preferences.

In this paper, we provide evidence against the above hypothesis. We design an algorithm to identify the most discriminative words and the character bigrams that are indicative of such words,

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 854–859,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

and perform two experiments to quantify their impact on the NLI task. The results of the first experiment demonstrate that the removal of a relatively small set of discriminative words from the training data significantly impairs the accuracy of a bigram-based classifier. The results of the second experiment reveal that the most indicative bigrams are quite similar across different language sets. We conclude that character bigrams are effective in determining L1 of the author because they reflect differences in L2 word usage that are unrelated to the phonology of L1.

2 Method

Tsur and Rappoport (2007) report that character bigrams are more effective for the NLI task than either unigrams or trigrams. We are interested in identifying the character bigrams that are indicative of the most discriminative words in order to quantify their impact on the bigram-based classifier.

We follow both Koppel et al. (2005) and Tsur and Rappoport (2007) in using a multi-class SVM classifier for the NLI task. The classifier computes a weight for each feature coupled with each L1 language by attempting to maximize the overall accuracy on the training set. For example, if we train the classifier using words as features, with values representing their frequency relative to the length of the document, the features corresponding to the word *China* might receive the following weights:

Arabic	Chinese	Hindi	Japanese	Telugu
-770	1720	-276	-254	-180

These weights indicate that the word provides strong positive evidence for Chinese as L1, as opposed to the other four languages.

We propose to quantify the importance of each word by converting its SVM feature weights into a single score using the following formula:

$$WordScore_i = \sqrt{\sum_{j=1}^{N} {w_{ij}}^2}$$

where N is the number of languages, and w_{ij} is the feature weight of word i in language j. The formula assigns higher scores to words with weights of high magnitude, either positive or negative. We use the Euclidean norm rather than the sum of raw weights because we are interested in the discriminative power of the words.

Algorithm 1 Computing the scores of words and bigrams in the data.

```
 1: create list of words in training data
 2: train SVM using words as features
 3: **for all** words i **do**
 4:     $WordScore_i = \sqrt{\sum_{j=1}^{N} {w_{ij}}^2}$
 5: **end for**
 6: sort words by WordScore
 7: NormValue = WordScore$_{200}$
 8: create list of 200 most frequent bigrams
 9: **for** bigrams k = 1 to 200 **do**
10:     $BigramScore_k = \prod_{k \in i} \frac{WordScore_i}{NormValue}$
11: **end for**
12: sort character bigrams by BigramScore
```

We normalize the word scores by dividing them by the score of the 200th word. Consequently, only the top 200 words have scores greater than or equal to 1.0. For our previous example, the 200^{th} word has a word score of 1493, while *China* has a word score of 1930, which is normalized to $1930/1493 = 1.29$. On the other hand, the 1000^{th} word gets a normalized score of 0.43.

In order to identify the bigrams that are indicative of the most discriminative words, we promote those that appear in the high-scoring words, and downgrade those that appear in the low-scoring words. Some bigrams that appear often in the high-scoring words may be very common. For example, the bigram `an` occurs in words like *Japan*, *German*, and *Italian*, but also by itself as a determiner, as an adjectival suffix, and as part of the conjunction *and*. Therefore, we calculate the importance score for each character bigram by multiplying the scores of each word in which the bigram occurs.

Algorithm 1 summarizes our method of identifying the discriminative words and indicative character bigrams. In line 2, we train an SVM on the words encountered in the training data. In lines 3 and 4, we assign the Euclidean norm of the weight vector of each word as its score. Starting in line 7, we determine which character bigrams are representative of high scoring words. In line 10, we calculate the bigram scores.

3 Experiments

In this section, we describe two experiments aimed at quantifying the importance of the discriminative words and the indicative character bigrams that are identified by Algorithm 1.

3.1 Data

We use two different NLI corpora. We follow the setup of Tsur and Rappoport (2007) by extracting two sets, denoted I1 and I2 (Table 1), from the International Corpus of Learner English (ICLE), Version 2 (Granger et al., 2009). Each set consists of 238 documents per language, randomly selected from the ICLE corpus. Each of the documents corresponds to a different author, and contains between 500 and 1000 words. We follow the methodology of the paper in performing 10-fold cross-validation on the sets of languages used by the authors.

For the development of the method described in Section 2, we used a different corpus, namely the TOEFL Non-Native English Corpus (Blanchard et al., 2013). It consists of essays written by native speakers of eleven languages, divided into three English proficiency levels. In order to maintain consistency with the ICLE sets, we extracted three sets of five languages apiece (Table 1), with each set including both related and unrelated languages: European languages that use Latin script (T1), non-European languages that use non-Latin scripts (T2), and a mixture of both types (T3). Each subcorpus was divided into a training set of 80%, and development and test sets of 10% each. The training sets are composed of approximately 700 documents per language, with an average length of 350 words per document. There are over 5000 word types per language, and over 1000 character bigrams in total. The test sets include approximately 90 documents per language. We report results on the test sets, after training on both the training and development sets.

3.2 Setup

We replicate the experiments of Tsur and Rappoport (2007) by limiting the features to the 200 most frequent character bigrams.[1] The feature values are set to the frequency of the character bigrams normalized by the length of the document. We use these feature vectors as input to the SVM-Multiclass classifier (Joachims, 1999). The results are shown in the *Baseline* column of Table 2.

ICLE:	
I1	Bulgarian Czech French Russian Spanish
I2	Czech Dutch Italian Russian Spanish
TOEFL:	
T1	French German Italian Spanish Turkish
T2	Arabic Chinese Hindi Japanese Telugu
T3	French German Japanese Korean Telugu

Table 1: The L1 language sets.

3.3 Discriminative Words

The objective of the first experiment is to quantify the influence of the most discriminative words on the accuracy of the bigram-based classifier. Using Algorithm 1, we identify the 100 most discriminative words, and remove them from the training data. The bigram counts are then recalculated, and the new 200 most frequent bigrams are used as features for the character-level SVM. Note that the number of the features in the classifier remains unchanged.

The results are shown in the *Discriminative Words* column of Table 2. We see a statistically significant drop in the accuracy of the classifier with respect to the baseline in all sets except T3. The words that are identified as the most discriminative include function words, punctuation, very common content words, and the toponymic terms. The 10 highest scoring words from T1 are: *indeed, often, statement, :* (colon), *question, instance, ...* (ellipsis), *opinion, conclude*, and *however*. In addition, *France*, *Turkey*, *Italian*, *Germany*, and *Italy* are all found among the top 70 words.

For comparison, we attempt to quantify the effect of removing the same number of randomly-selected words from the training data. Specifically, we discard all tokens that correspond to 100 word types that have the same or slightly higher frequency as the discriminative words. The results are shown in the *Random Words* column of Table 2. The decrease is much smaller for I1, I2, and T1, while the accuracy actually increases for T2 and T3. This illustrates the impact that the most discriminative words have on the bigram-based classifier beyond simple reduction in the amount of the training data.

[1] Our development experiments suggest that using the full set of bigrams results in a higher accuracy of a bigram-based classifier. However, we limit the set of features to the 200 most frequent bigrams for the sake of consistency with previous work.

Set	Baseline	Random Words	Discriminative Words	Random Bigrams	Indicative Bigrams
I1	67.5	−0.2	−3.6	−1.0	−2.2
I2	66.9	−2.5	−5.5	−0.7	−2.8
T1	60.7	−3.3	−7.7	−2.5	−3.9
T2	60.6	+0.5	−3.8	−1.1	−5.9
T3	62.2	+0.3	−0.0	−0.5	−4.1

Table 2: The impact of subsets of word types and bigram features on the accuracy of a bigram-based NLI classifier.

3.4 Indicative Bigrams

Using Algorithm 1, we identify the top 20 character bigrams, and replace them with randomly selected bigrams. The results of this experiment are reported in the *Indicative Bigrams* column of Table 2. It is to be expected that the replacement of any 20 of the top bigrams with 20 less useful bigrams will result in some drop in accuracy, regardless of which bigrams are chosen for replacement. For comparison, the *Random Bigrams* column of Table 2 shows the mean accuracy over 100 trials obtained when 20 bigrams randomly selected from the set of 200 bigrams are replaced with random bigrams from outside of the set.

The results indicate that our algorithm indeed identifies 20 bigrams that are on average more important than the other 180 bigrams. What is really striking is that the sets of 20 indicative character bigrams overlap substantially across different sets. Table 3 shows 17 bigrams that are common across the three TOEFL corpora, ordered by their score, together with some of the highly scored words in which they occur. Four of the bigrams consist of punctuation marks and a space.[2] The remaining bigrams indicate function words, toponymic terms like *Germany*, and frequent content words like *take* and *new*.

The situation is similar in the ICLE sets, where likewise 17 out of 20 bigrams are common. The inter-fold overlap is even greater, with 19 out of 20 bigrams appearing in each of the 10 folds. In particular, the bigrams `fr` and `bu` can be traced to both the function words *from* and *but*, and the presence of French and Bulgarian in I1. However, the fact that the two bigrams are also on the list for the I2 set, which does not include these languages, suggests that their importance is mostly due to the function words.

Bigram	Words
_,	
,_	
_.	
._	
u_	you Telugu
f_	of
ny	any many Germany
yo	you your
w_	now how
i_	I
_y	you your
ew	new knew
kn	know knew
ey	they Turkey
wh	what why where *etc.*
of	of
ak	make take

Table 3: The most indicative character bigrams in the TOEFL corpus (sorted by score).

3.5 Discussion

In the first experiment, we showed that the removal of the 100 most discriminative words from the training data results in a significant drop in the accuracy of the classifier *that is based exclusively on character bigrams*. If the hypothesis of Tsur and Rappoport (2007) was true, this should not be the case, as the phonology of L1 would influence the choice of words across the lexicon.

In the second experiment, we found that the majority of the most indicative character bigrams *are shared among different language sets*. The bigrams appear to reflect primarily high-frequency function words. If the hypothesis was true, this

[2] It appears that only the relatively low frequency of most of the punctuation bigrams prevents them from dominating the sets of the indicative bigrams. When using all bigrams instead of the top 200, the majority of the indicative bigrams contain punctuation.

should not be the case, as the diverse L1 phonologies would induce different sets of bigrams. In fact, the highest scoring bigrams reflect punctuation patterns, which have little to do with word choice.

4 Conclusion

We have provided experimental evidence against the hypothesis that the phonology of L1 strongly affects the choice of words in L2. We showed that a small set of high-frequency function words have disproportionate influence on the accuracy of a bigram-based NLI classifier, and that the majority of the indicative bigrams appear to be independent of L1. This suggests an alternative explanation of the effectiveness of a bigram-based classifier in identifying the native language of a writer — that the character bigrams simply mirror differences in the word usage rather than the phonology of L1.

Our explanation concurs with the findings of Daland (2013) that unigram frequency differences in certain types of phonological segments between child-directed and adult-directed speech are due to a small number of word types, such as *you*, *what*, and *want*, rather than to any general phonological preferences. He argues that the relative frequency of sounds in speech is driven by the relative frequency of words. In a similar vein, Koppel et al. (2005) see the usefulness of character n-grams as "simply an artifact of variable usage of particular words, which in turn might be the result of different thematic preferences," or as a reflection of the L1 orthography.

We conclude by noting that our experimental results do not imply that the phonology of L1 has absolutely no influence on L2 writing. Rather, they show that the evidence from the Native Language Identification task has so far been inconclusive in this regard.

Acknowledgments

We thank the participants and the organizers of the shared task on NLI at the BEA8 workshop for sharing their reflections on the task. We also thank an anonymous reviewer for pointing out the study of Daland (2013).

This research was supported by the Natural Sciences and Engineering Research Council of Canada and the Alberta Innovates Technology Futures.

References

Amjad Abu-Jbara, Rahul Jha, Eric Morley, and Dragomir Radev. 2013. Experimental results on the native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–88.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A Corpus of Non-Native English. Technical report, Educational Testing Service.

Julian Brooke and Graeme Hirst. 2013. Using other learner corpora in the 2013 NLI shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 188–196.

Serhiy Bykh, Sowmya Vajjala, Julia Krivanek, and Detmar Meurers. 2013. Combining shallow and linguistically motivated features in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 197–206.

Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2013. Linguistic profiling based on general–purpose features and native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–215.

Robert Daland. 2013. Variation in the input: a case study of manner class frequencies. *Journal of Child Language*, 40(5):1091–1122.

Vidas Daudaravicius. 2013. VTEX system description for the NLI 2013 shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 89–95.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with TF-IDF weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 216–223.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2013. Feature space selection and combination for native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 96–100.

Sylvaine Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. INTERNATIONAL CORPUS OF LEARNER ENGLISH: VERSION 2.

John Henderson, Guido Zarrella, Craig Pfeifer, and John D. Burger. 2013. Discriminating non-native English with 350 words. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 101–110.

Barbora Hladka, Martin Holub, and Vincent Kriz. 2013. Feature engineering in the NLI shared task 2013: Charles University submission report. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 232–241.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118.

Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. In *Advances in kernel methods*, pages 169–184. MIT Press.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628, Chicago, IL. ACM.

Shibamouli Lahiri and Rada Mihalcea. 2013. Using n-gram and word network features for native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 251–259.

Baoli Li. 2013. Recognizing English learners. native language from their writings. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 119–123.

André Lynum. 2013. Native language identification using large scale lexical features. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 266–269.

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. 2013. NLI shared task 2013: MQ submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133.

Tomoya Mizumoto, Yuta Hayashibe, Keisuke Sakaguchi, Mamoru Komachi, and Yuji Matsumoto. 2013. NAIST at the NLI 2013 shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 134–139.

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Lei Yao, and Grzegorz Kondrak. 2013. Cognate and misspelling features for natural language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 140–145.

Marius Popescu and Radu Tudor Ionescu. 2013. The story of the characters, the DNA and the native language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 270–278.

Ben Swanson. 2013. Exploring syntactic representations for native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 146–151.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A Report on the First Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*.

Oren Tsur and Ari Rappoport. 2007. Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16, Prague, Czech Republic.

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. 2013. Identifying the L1 of non-native writers: the CMU-Haifa system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 279–287.