Experimental Results on the Native Language Identification Shared Task

Amjad Abu-Jbara, Rahul Jha, Eric Morley, Dragomir Radev
Department of EECS
University of Michigan
Ann Arbor, MI, USA
[amjbara, rahuljha, eamorley, radev]@umich.edu

Abstract

We present a system for automatically identifying the native language of a writer. We experiment with a large set of features and train them on a corpus of 9,900 essays written in English by speakers of 11 different languages. our system achieved an accuracy of 43% on the test data, improved to 63% with improved feature normalization. In this paper, we present the features used in our system, describe our experiments and provide an analysis of our results.

1 Introduction

The task of Native Language Identification (NLI) is the task of identifying the native language of a writer or a speaker by analyzing their writing in English. Previous work in this area shows that there are several linguistic cues that can be used to do such identification. Based on their native language, different speakers tend to make different kinds of errors pertaining to spelling, punctuation, and grammar (Garfield, 1964; Wong and Dras, 2009; Kochmar, 2011). We describe the complete set of features we considered in Section 4. We evaluate different combinations of these features, and different ways of normalizing them in Section 5.

There are many possible applications for an NLI system, as noted by Kochmar (2011): finding the origins of anonymous text; error correction in various tasks including speech recognition, part-of-speech tagging, and parsing; and in the field of second language acquisition for identifying learner difficulties. We are most interested in statistical approaches to this problem because it may point towards fruitful avenues of research in language and sound transfer, which are how people apply knowledge of their native language, and its phonology and orthography, respectively, to a second language. For example, Tsur and Rappoport (2007) found that character bigrams are quite useful for NLI, which led them to suggest that second language learners' word choice may in part be driven by their native languages. Analysis of such language and sound translation patterns might be useful in understanding the process of language acquisition in humans.

2 Previous Work

The work presented in this paper was done as part of the NLI shared task (Tetreault et al., 2013), which is the first time this problem has been the subject of a shared task. However, several researchers have investigated NLI and similar problems. Authorship attribution, a related problem, has been well studied in the literature, starting from the seminal work on disputed Federalist Papers by Mosteller and Wallace (1964). The goal of authorship attribution is to assign a text to one author from a candidate set

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 82–88,
Atlanta, Georgia, June 13 2013.

of authors. This technique has many applications, and has recently been used to investigate terrorist communication (Abbasi and Chen, 2005) and digital crime (Chaski, 2005). The goal of NLI somewhat similar to authorship attribution, in that NLI attempts to distinguish between candidate communities of people who share a common cultural and linguistic background, while authorship attribution distinguishes between candidate individuals.

In the earliest treatment of this problem, Koppel et al. (2005) used stylistic text features to identify the native language of an author. They used features based on function words, character n-grams and errors and idiosyncrasies such as spelling errors and non-standard syntactic constructions. They experimented on a dataset with essays written by non-native English speakers from five countries, Russia, Czech Republic, Bulgaria, France and Spain, with 258 instances from each dataset. They trained a multi-class SVM model using the above features and reported 10-fold cross validation accuracy of 80.2%.

Tsur and Rappoport (2007) studied the problem of NLI with a focus on *language transfer*, i.e. how a seaker's native language affects the way in which they acquire a second language, an important area in Second Language Acquisition research. Their feature analysis showed that character bigrams alone can lead to a classification accuracy of about 66% in a 5-class task. They concluded that the choice of words people make when writing in a second language is highly influenced by the phonology of their native language.

Wong and Dras (2009) studied syntactic errors derived from contrastive analysis as features for NLI. They used the five languages from Koppel et al. (2008) along with Chinese and Japanese, but did not find an improvement in classification accuracy by adding error features based on contrastive analysis. Later, Wong and Dras (2011) studied a more general set of syntactic features and showed that adding these features improved the accuracy significantly. They also investigated classification models based on LDA (Wong et al., 2011), but did not find them to be useful overall. They did, however, notice that some of the topics were capturing information that would be useful for identifying particular native languages. They also proposed the use of adaptor grammars (Johnson et al., 2007), which are a generalization of probabilistic context-free grammars, to capture collocational pairings. In a later paper, Wong et al. explored the use of adapter grammars in detail (Wong et al., 2012) and showed that an extension of adaptor grammars to discover collocations beyond lexical words can produce features useful for the NLI task.

3 Dataset

The experiments for this paper were performed using the TOEFL11 dataset (Blanchard et al., 2013) provided as part of the shared task. The dataset contains essays written in English from native speakers of 11 languages (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish). The corpus contains 12,099 essays per language sampled evenly from 8 prompts or topics. This dataset was designed specifically to support the task of NLI and addresses some of the shortcomings of earlier datasets used for research in this area. Specifically, the dataset has been carefully selected in order to maintain consistency in topic distributions, character encodings and annotations across the essays from different native languages. The data was split into three data sets: a training set comprising 9,900 essays, a development set comprising 1,100 essays, and a test set comprising 1,100 essays.

4 Approach

We addressed the problem as a supervised, multi-class classification task. We trained a Support Vector Machine (SVM) classifier on a set of lexical, syntactic and dependency features extracted from the training data. We computed the minimum and maximum values for each of the features and normalized the values by the range (max - min). Here we describe the features in turn.

Character and Word N-grams Tsur and Rappoport (2007) found that character bigrams were useful for NLI, and they suggested that this may be due to the writer's native language influencing their choice of words. To reflect this, we compute features using both characters and word N-grams. For characters, we consider 2,3 and 4-grams, with padding characters at the beginning and end of each sentence. The features are generated over the entire training data, i.e., every n-gram occurring in the training data is used as a feature. Similarly, we create features with 1,2 and 3-grams of words. Each word n-gram is used as a separate feature. We explore both binary features for each character or word n-gram, as well as normalized count features.

Part-Of-Speech N-grams Several investigations, for example those conducted by Kochmar (2011) and Wong and Dras (2011), have found that part-of-speech tags can be useful for NLI. Therefore we include part-of-speech (POS) n-grams as features. We parse the sentences with the Stanford Parser (Klein and Manning, 2003) and extract the POS tags. We use binary features describing the presence or absence of POS bigrams in a document, as well as numerical features describing their relative frequency in a document.

Function Words Koppel et al. (2005) found that function words can help identify someone's native language. To this end, we include a categorical feature for the presence of function words that are included in list of 321 function words.

Use of punctuation Based on our experience with speakers of native languages, as well as Kochmar's (2011) observations of written English produced by Germanic and Romance language speakers, we suspect that speakers of different native languages use punctuation in different ways, presumably based on the punctuation patterns in their native language. For example, comma placement differs between German and English, and neither Chinese nor Japanese requires a full stop at the end of every sentence. To capture these kinds of patterns, we create two features for each essay: the number of punctuation marks used per sentence, and the number of punctuation marks used per word.

Number of Unique Stems Speakers of different native languages might differ in the amount of vocabulary they use when communicating in English. We capture this by counting the number of unique stems in each essay and using this as an additional feature. The hypothesis here is that depending on the similarity of the native language with English, the presence of common words, and other cultural cues, people with different native language might have access to different amounts of vocabulary.

Misuse of Articles We count instances in which the number of an article is inconsistent with the associated noun. To do so, we fist identify all the *det* dependency relations in the essay. We then compute the ratio of *det* relations between 'a' or 'an' and a plural noun (NNS), to all *det* relations. We also count the ratio of *det* relations between 'a' or 'an' and an uncountable noun, to all *det* relations. We do this using a list of 288 uncountable nouns.[1]

Capitalization The writing systems of some languages in the data set, for example Telugu, do not include capitalization. Furthermore, other languages may use capitalization quite differently from English, for example German, in which all nouns are capitalized, and French, in which nationalities are not. Character capitalization mistakes may be common in the text written by the speakers of such languages. We compute the ratio of words with at least two letters that are written in all caps to identify excessive capitalization. We also count the relative frequency of capitalized words that appear in the middle of a sentence that are not tagged as proper nouns by the part of speech tagger.

Tense and Aspect Frequency Verbal tense and aspect systems vary widely between languages. English has obligatory tense (past, present, future) and

[1] http://www.englishclub.com/vocabulary/nouns-uncountable-list.htm

aspect (imperfect, perfect, progressive) marking on verbs. Other languages, for example French, may require verbs to be marked for tense, but not aspect. Still other languages, for example Chinese, may use adverbials and temporal phrases to communicate temporal and aspectual information. To attempt to capture some of the ways learners of English may be influenced by their native language's system of tense and aspect, we compute two features. First, we compute the relative frequency of each tense and aspect in the article from the counts of each verb POS tags (ex. VB, VBD, VBG). We also compute the percentage of sentences that contain verbs of different tenses or aspect, again using the verb POS tags.

Missing Punctuation We compute the relative frequency of sentences that include an introductory phrase (e.g. however, furthermore, moreover) that is not followed by a comma. We also count the relative frequency of sentences that start with a subordinating conjunction (e.g. sentences starting with if, after, before, when, even though, etc.), but do not contain a comma.

Average Number of Syllables We count the number of syllables per word and the ratio of words with three or more syllables. To count the number of syllables in a word, we used a perl module that estimates the number of syllables by applying a set of hand-crafted rules.[2].

Arc Length We calculate several features pertaining to dependency arc length and direction. We parse each sentence separately, using the Stanford Dependency Parser, and then compute a single value for each of these features for each document. First, we simply compute the percentage of arcs that point left or right (PCTARCL, PCTARCR). We also compute the minumum, maximum, and mean dependency arc length, ignoring arc direction. We also compute similar features for typed dependencies: the minimum, maximum, and mean dependency arc length for each typed dependency; and the percentage of arcs for each typed dependency that go to the left or right.

Downtoners and Intensifiers We compute three features to describe the use of downtoners, and three for intensifiers in each document. First, we count the number of downtoners or intensifiers in a given document.[3] We normalize this count by the number of tokens, types, and sentences in the document to yield the three features capturing the use of downtoners or intensifiers.

Production Rules We compute a set of features to describe the relative frequency of production rules in a given document. First, we parse each sentence using the Stanford Parser, using the default English PCFG (Klein and Manning, 2003). We then count all non-terminal production rules in a given document, and report the relative frequency of each production rule in that document.

Subject Agreement We count the number of sentences in which there appears to be a mistake in subject agreement. To do this, we first identify *nsubj* and *nsubjpass* dependency relationships. Of these dependencies, we count ones meeting the following criteria as mistakes: a third person singular present tense verb with a nominal that is not third person singular, and a third person singular subject with a present tense verb not marked as third person singular. We then normalize the count of errors by the total number of nsubj and nsubj pass dependencies in the document, and the number of sentences in the document to produce two features.

Words per Sentence We compute both the number of tokens per line and the number of types per

[2]http://search.cpan.org/dist/Lingua-EN-Syllable/Syllable.pm

[3]The words we count as downtoners are: 'almost', 'alot', 'a lot', 'barely', 'a bit', 'fairly', 'hardly', 'just', 'kind of', 'least', 'less', 'merely', 'mildly', 'nearly', 'only', 'partially', 'partly', 'practically', 'rather', 'scarcely', 'sort of', 'slightly', and 'somewhat'. The intensifiers are: 'a good deal', 'a great deal', 'absolutely', 'altogether', 'completely','enormously', 'entirely', 'extremely', 'fully', 'greatly', 'highly', 'intensely', 'more', 'most', 'perfectly', 'quite', 'really', 'so', 'strongly', 'super', 'thoroughly', 'too', 'totally', 'utterly', and 'very'.

line.

Topic Scores We construct an unsupervised topic model for all of the documents using Mallet (McCallum, 2002) with 100 topics, dirichlet hyperparameter reestimation every 10 rounds, and all other options set to default values. We then use the topic weights as features.

Passive Constructions We count the number of times an author uses passive constructions by counting the number of *nsubjpass* dependencies in each document. We normalize this count in two ways to produce two different features: dividing by the number of sentences, and dividing by the total number of nsubj and nsubjpass dependencies.

5 Experiments and Results

We used weka (Hall et al., 2009) and libsvm (Chang and Lin, 2011) to run the experiments. The classification was done using an SVM classifier. We experimented with different SVM kernels and different values for the cost parameter. The best performance was achieved with a linear kernel and $cost = 0.001$. We trained the model using the combination of the training and the development sets. We submitted the output of the system to the NLI shared task workshop. Our system achieved 43.3% accuracy. Table 1 shows the confusion matrix and the precision, recall, and F-measure for each language. After the NLI submission deadline, we noticed that we our system was not handling the normalization of the features properly which resulted in the poor performance. After fixing the problem, our system achieved 63% accuracy on both test data and 10-fold cross validation on the entire data.

6 Analysis

We did feature analysis on the training and development data sets using the Chi-squared test. Our feature analysis shows that the most important features for the classifier were topic models, character n-grams of all orders, word unigrams and bigrams, POS bigrams, capitalization features, function words, production rules, and arc length. These results are consistent with those presented in previous work done on this task.

Looking at the confusion matrix in Figure 1, we see that Korean and Japanese were the most commonly confused pair of languages. Hindi and Telugu, two languages from the Indian subcontinent, were also often confused. To analyze this further, we did another experiment by training just a binary classifier on Korean and Japanese using the exact same feature set as earlier. We achieved a 10-fold cross validation accuracy of 83.3% on this classification task. Thus, given just these two languages, we were able to obtain high classification accuracy. This suggests that a potentially fruitful strategy for NLI systems might be to fuse often-confused pairs, such as Korean/Japanese and Hindi/Telugu, into singleton classes for the initial run, and then run a second classifier to do a more fine grained classification within these higher level classes.

When doing feature analysis for these two languages, we found that the character bigrams representing the country names were some of the top features used for classification. For example "Kor" occurred as a trigram frequently in essays from native language speakers of Korean. Based on this, we designed a small experiment where we created features corresponding to the country name associated with each native language, e.g., "Korea", "China", "India", "France", etc. For Arabic, we used a list of 22 countries where Arabic is spoken. Just using this feature, we obtained a 10-fold cross validation accuracy of 21.3% on the development set. This suggests that in certain genres, one may be able to leverage information conveying geographical and demographic attributes for NLI.

7 Conclusion

In this paper, we presented a supervised system for the task of Native Language Identification. We describe and motivate several features for this task and report results of supervised classification using these features on a test data set consisting of 11 lan-

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR	Precision	Recall	F-measure
ARA	41	7	8	3	6	2	3	5	10	7	8	44.6%	41.0%	42.7%
CHI	6	38	5	2	2	8	15	8	3	3	10	40.0%	38.0%	39.0%
FRE	8	6	43	8	1	14	2	4	6	1	7	39.1%	43.0%	41.0%
GER	3	3	10	49	4	9	1	7	6	0	8	54.4%	49.0%	51.6%
HIN	5	2	6	9	34	0	3	1	3	32	5	47.9%	34.0%	39.8%
ITA	5	3	10	5	1	52	2	1	17	0	4	46.0%	52.0%	48.8%
JPN	3	11	0	1	1	3	49	26	1	1	4	37.4%	49.0%	42.4%
KOR	2	6	6	1	1	2	35	40	1	1	5	38.1%	40.0%	39.0%
SPA	4	6	14	1	1	17	6	2	38	0	11	40.9%	38.0%	39.4%
TEL	9	7	3	4	18	2	2	2	2	48	3	51.1%	48.0%	49.5%
TUR	6	6	5	7	2	4	13	9	6	1	41	38.7%	41.0%	39.8%
Accuracy = 43.0%														

Table 1: The results of our original submission to the NLI shared task on the test set. These results reflect the performance of the system that does not normalize the features properly

guages provided as part of the NLI shared task. We found that our classifier often confused two pairs of languages that are spoken near one another, but are linguistically unrelated: Hindi/Telugu and Korean/Japanese. We found that we could obtain high classification accuracy on these two pairs of languages using a binary classifier trained on just these pairs. During our feature analysis, we also found that certain features that happened to convey geographical and demographic information were also informative for this task.

References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, September.

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

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Carole E. Chaski. 2005. Who's at the keyboard: Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4:2005.

Eugene Garfield. 1964. Can citation indexing be automated?

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, 19:641.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Ekaterina Kochmar. 2011. *Identification of a Writer's Native Langauge by Error Analysis*. Ph.D. thesis.

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.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2008. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist Papers*. Addison-Wesley, Reading, Mass.

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, Atlanta, GA, USA, June. Association for Computational Linguistics.

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*, CACLA '07, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive Analysis and Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia, December.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting Parse Structures for Native Language Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2011. Topic Modeling for Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124, Canberra, Australia, December.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring Adaptor Grammars for Native Language Identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 699–709, Jeju Island, Korea, July. Association for Computational Linguistics.