”All Your Base Are Belong To Us”: Identifying English Texts Written by Non-Native Speakers

Jonathan Hung
Stanford University
hungj@stanford.edu

Joanna Kim
Stanford University
joannak@stanford.edu

I. Introduction

English is one of the most prevalent languages worldwide, third only to Mandarin Chinese and Spanish. With millions of people learning English as a second language, it is a worthwhile endeavor to improve their experience by creating learning tools customized to them. For our project, we created a classifier that can take raw English text and identify the writer as either an English-native speaker or an English learner. Furthermore, we expanded the classifier so that it could identify the writer's native language given a raw text file. This classifier could become an important basis for a learning tool (i.e. an editing tool that corrects and gives suggestions specialized to the writer's native language) that can help ESL students gain a better grasp of the English language.

II. Data

Our data is in the format of raw text files that have been written by both English native speakers and non-native speakers. We've drawn our data from one online corpus called the ICNALE [2], the International Corpus Network of Asian Learners of English, which has 5,200 short answer essays written by ESL students from countries ranging from China to Singapore to Pakistan for a total of 10 different Asian countries. It also contains a smaller set of essays written by English native speakers from the United States, United Kingdom, Australia, and New Zealand. The exact counts of essays per country are enumerated in Table I. These essays are responses to two specific questions (the first talking about part-time jobs in college, and the second about smoking in restaurants). All the essays are short in length with only approximately 5 to 10 sentences each.

Research shows that many of the errors that ESL students make can be correlated to the structure of their own native language [4]. Thus, we've decided to divide and conquer and focus our attention on errors and features specific to a defined sector of languages that have similar structures. Eventually, we would like to expand to ESL learners based in other languages, like the Romantic or Germanic languages.

Country	Count
English-Speaking (ENS)	400
China (CHN)	800
Hong Kong (HKG)	200
Indonesia (IDN)	400
Japan (JPN)	800
Korea (KOR)	600
Pakistan (PAK)	400
Philippines (PHL)	400
Singapore (SIN)	400
Thailand (THA)	800
Taiwan (TWN)	400

TABLE I: Number of Essays per Country

III. Feature Selection

For our three models (Logistic Regression, Naive Bayes, and a Markov Model with n-grams), we used different sets of features because we were interested how they would perform given different features. We used three different types of features, frequently used in Native Language Identification (NLI): features by grammatical cues, features by frequency of words, and features by parts-of-speech n-grams.

For Logistic Regression, we focused on features that would detect grammatical/syntatical cues common among non-native speakers. We have two sets of syntactical features: one set that required minimum manipulation of the text to extract the features while the other required syntactic parsing to extract grammatical features. Our first set has three features (sentence length, misspellings, and the repetition of words). For our second set, we used the Stanford Parts of Speech Tagger to label words by their part of speech (e.g. noun, verb, adverb, etc.) and then created twelve different features by parts of speech. We chose these features because, though

not the most common, they are relatively common cues of non-native speakers and they can be easily be derived from raw text. A good future extension might be to expand the feature list to features that are statistically more common among non-native speakers (errors with articles and prepositions, the presence of fragments and run-on sentences, awkward or missing diction) but that require the use of the Stanford Parser to understand their grammatical context [4].

For Naive Bayes, we used features at the granularity level of words - for each word, we count its frequency and use these as our features. We omitted words that occurred in total fewer than three times, since the sparseness of these words would not add any predictive power, and it also drastically reduced our data size.

For n-gram Markov Model, we labeled each word with the Stanford Parts of Speech Tagger and created features based off of strings of 2-grams.

IV. Models and Results

A. Non-Native English Classification

Our first two models (Logistic Regression and Naive Bayes) focus on binary classification. Given a raw text of English, they aim at classifying if the writer is a native or non-native speaker.

1) Logistic Regression: We used our logistic regression model to test all the syntactical features discussed in the Feature Selection section. Using a 15 dimensional feature space, we used the following hypothesis in our regression:

$$h_\theta(x) = 1\left\{\frac{1}{1+\exp(-\theta^T x)} \geq 0.5\right\}$$

We trained θ with Batch Stochastic Gradient and a learning rate α of 0.00001 over a training set of size 1080. We then tested θ over a test set of size 120. Both sets were a mixture of Asian-language native speakers and English-native speakers. For our results, we define error as follows:

$$\epsilon = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} \neq h_\theta(x^{(i)})\}.$$

2) Naive Bayes: With the hypothesis that English native speakers might be inclined to use words that non-native speakers wouldn't and vice-versa, we implemented a Naive Bayes algorithm that finds the probabilities that a specific essay would be written given that the writer is an English native speaker or a learner. Then, the algorithm categorizes the essay by the higher probability. We used a multinomial event model which is suited for text classification and Laplace smoothing to account for words that were not in the data.

We trained our algorithm on a set of 1080 samples and tested on a separate set of 120 samples. Both sets were mixed with Asian-Language native and English-native writing samples. Similarly to how we calculated error for our logistic regression model, we calculated error as follows:

$$\epsilon = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} \neq h_\theta(x^{(i)})\}.$$

B. Language Classification

We next turn to the problem of classifying a text by the country of origin of the writer. For example, can we identify that a text comes from a Chinese writer versus from Korean writer?

To do this, we implement a Markov model using n-grams as our states. For our results, we use $n = 2$. The methodology is as follows:

We first convert each of the essays in our training data to a list of parts of speech using Stanford's parts of speech tagger [5]. For example, the sentence "This is a paper" would be converted to (determiner, third person verb, determiner, singular noun). We then take consecutive 2-sequences of parts of speech, and count the frequency of each 2-sequence in all of the training essays for a language of origin. Thus, each language has its own model of parts of speech frequencies. Then, for each essay in our test data, we find the likelihood of the sequence of parts of speech from that essay appearing in each language based on our models. The prediction is the language that results in the highest likelihood.

V. Results

A. Non-Native English Classification

1) Logistic Regression: Our error for our training set was 17.4% while our error for test set was 15.8%. Our test set converged after 11,841 iterations, a reasonable amount given our tiny learning rate. Our model did surprisingly well, considering that the features chosen were not the best indicators of non-native speakers.

	NS	NNS
NS	28	16
NNS	3	73

TABLE II: Confusion matrix for logistic regression

Table II shows our results for logistic regression.

2) Naive Bayes: For the training set error, we found a very accurate 0.09% error rate, and for our testing set error, we have a 1.6% error rate, with the confusion matrix shown in Table III.

	NS	NNS
NS	39	1
NNS	1	79

TABLE III: Confusion matrix for Naive Bayes

B. Language Classification

Using the Markov model described previously, we achieved a training error of 24.3% and a test error of 34.7%, with the confusion matrix shown in Table IV.

In addition, we plotted accuracies and recalls for each language in Figure 1 and Figure 2.

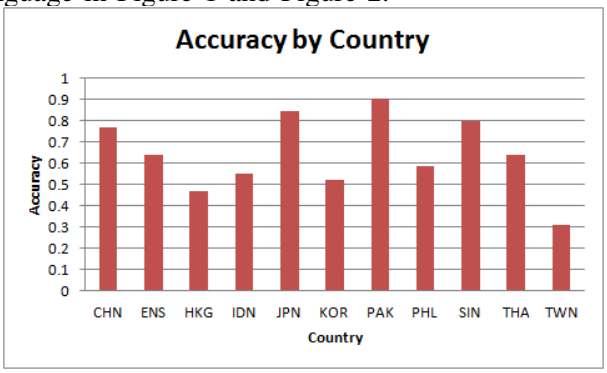

Fig. 1: Accuracy by Country

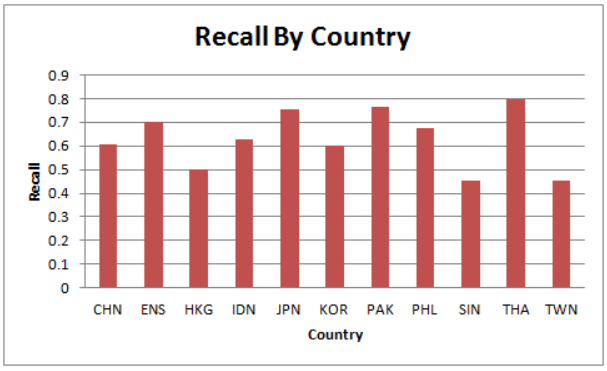

Fig. 2: Recall by Country

VI. Discussion

We were happy with our results from all three models. Logistic regression worked well considering some of our features were basic features of text. Note that our non-native speaker error was $3/76 = 3.9\%$ while our native speaker error was $16/44 = 36.4\%$. The discrepancy is likely due to the fact that we had much more non-native speaker data, so we were able to build a better model for it.

We were even more successful with Naive Bayes. Our results mirrored the research that implied that semantic cues - such as the diction that our Naive Bayes algorithm measures - are excellent indicators of learners' writing [1].

The results from the Markov model were particularly interesting. Using a basic metric such as consecutive parts of speech, we were able to build reasonably good models for what the text from a writer of a certain country looks like. It is also notable where some of the misclassifications occur. For example, 14 Korean essays were misclassified as Japanese. This can potentially explained by the fact that these languages are influenced by Chinese, so their writers might have similar writing patterns. Also, many Taiwanese essays (10) were misclassified as Chinese, which makes sense since people in Taiwan speak Mandarin.

The accuracies and recalls by country also reflect some of these patterns; for example, Taiwanese recall is very low, considering that many of them were classified as Chinese, a country that speaks the same (Mandarin) or similar (Taiwanese) language. In addition, Hong Kong had low accuracy and recall, possibly due to the lack of data in relation to the other countries, so we could not build as good a model for it.

Finally, we see both high accuracy and recall for Pakistan. This is possibly due to the fact that Pakistan is unlike the other languages in grammatical structure and diction, and thus, it was less likely to be confused with other languages and more likely to be accurately labelled.

VII. Future Plans

There are a number of ways to expand and improve on our current models through a number of ways. For one, we would like to put our models through more rigorous tests that have a greater variety in both the native languages of the writers and the topics of the essays. We would also like to add more features that require more grammatical parsing but are very common among non-native speakers. Additionally in the near future, we would like to add another algorithm that uses character n-grams and string kernels to categorize texts by their writers native language. N-grams by character would allow the classifier to abstract away language features like parts-of-speech, diction, and syntax. Thus, the classifier could easily used for languages other than English since it would not be based on linguistic structure.

	CHN	ENS	HKG	IDN	JPN	KOR	PAK	PHL	SIN	THA	TWN
CHN	53	0	0	2	2	3	2	0	1	3	3
ENS	0	21	0	0	1	0	0	2	9	0	0
HKG	5	0	8	0	0	1	0	1	2	0	0
IDN	3	1	1	22	1	3	0	1	1	5	2
JPN	2	1	0	0	71	5	0	0	0	1	4
KOR	4	2	2	2	14	35	0	1	1	4	2
PAK	0	0	0	0	0	1	29	1	1	0	0
PHL	4	1	0	2	0	1	3	27	7	1	0
SIN	1	2	1	0	0	0	0	2	24	0	0
THA	5	0	2	4	5	5	4	4	0	62	6
TWN	10	2	2	3	0	4	0	1	7	2	14

TABLE IV: Confusion matrix for Markov Model

REFERENCES

[1] Heilman, Michael J., Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskanazi. *Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts*. (n.d.): n. pag. Web.

[2] Ishikawa, Dr. Shin'ichiro. *ICNALE: The International Corpus Network of Asian Learners of English.* N.p., n.d. Web. 14 Nov. 2014.

[3] *Korean Learner Corpus Blog.* : Korean Learner Corpora. N.p., n.d. Web. 14 Nov. 2014.

[4] Leacock, Claudia, Martin Chodorow, Michael Gamon, and Joel Tetreault. *Automated Grammatical Error Detection for Language Learners.* N.p.: n.p., n.d. Web.

[5] http://nlp.stanford.edu/software/tagger.shtml