Identifying the native language of non-native English writers

Chris Dyer, Manaal Faruqui, Nathan Schneider, Yulia Tsvetkov, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA
Noam Ordan, Naama Twitto, Shuly Wintner, Department of Computer Science,University of Haifa, Haifa, Israel

Abstract

We show that it is possible to identify, with high accuracy, the native language of English test takers from the content of the essays they write. We use standard machine-learning text classification techniques to predict the most probable native language from a set of 11 possibilities: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish.

Motivation

Our motivation is both to improve the accuracy of classification and to be able to interpret the characteristics of the language produced by speakers of different native languages. Such a system can be useful for a number of applications:

- Tutoring applications that can identify the native language of the writer can offer better targeted feedback to language learners about their errors, since many errors are native language-specific.
- Forensic linguistic applications often need to determine the native language of authors.

Datasets

The dataset consists of 12,100 English TOEFL essays written by speakers of 11 different languages, distributed as part of the First Native Language Identification Shared Task. Each essay is labelled with the author's English language proficiency level (high, medium, or low) and an identification (1 to 8) of the essay prompt.

Data	# essays
Training	9,900
Development	1,100
Testing	1,100

Number of essays per dataset.

	Low	Medium	High
Essays	1,069	5,366	3,456
Tokens	245,130	1,819,407	1,388,260
Types	13,110	37,393	28,329

Corpus size per author's proficiency level.

Main Features

This system was developed to classify each essay to one of 11-class languages. For our classification model, we use the `creg` regression modelling framework (`https://github.com/redpony/creg`).We define a large arsenal of features. The following four feature types are used as the baseline features:

Part-of-speech (POS) n-grams Features were extracted to count every POS 1-, 2-, and 3-gram in each document.

FreqChar We experimented with character n-grams. To reduce the number of parameters, we retain only those character n-grams that are observed more than 5 times in the training corpus. The value of n ranges from 1 to 4.

CharPrompt Conjunction of the character n-gram features defined above with the prompt ID

Brown clusters Using the Brown clusters algorithm we clustered 8 billion words of English into 600 classes. Then we included log counts of all 4-grams of Brown clusters that occurred at least 100 times in the training data.

Additional Features

To the basic set of features we add more specific, linguistically-motivated features such as:

PsvRatio The proportion of passive verbs out of all verbs.

DocLen Document length in tokens.

Punct Log counts of each punctuation mark.

Misspell Spelling correction edits. Features included substitutions, deletions, insertions and joinings (*alot*→*a lot*) that were made by the author of the essay. High-weight features include: `ARA:DEL<e>`, `ARA:INS<e>`; `GER:SUBST<z>/<y>`; `JPN:SUBST<l>/<r>`, `JPN:SUBST<r>/<l>`; `SPA:INS<s>`.

CxtFxn Contextual function words. We define pairs consisting of a function word, along with the POS tag of its adjacent word. We also define 3-grams consisting of one or two function words and the POS tag of the third word in the 3-gram.

Results

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	80	0	2	1	3	4	1	0	4	2	3
CHI	3	80	0	1	1	0	6	7	1	0	1
FRE	2	2	81	5	1	2	1	0	3	0	3
GER	1	1	1	93	0	0	0	1	1	0	2
HIN	2	0	0	1	77	1	0	1	5	9	4
ITA	2	0	3	1	1	87	1	0	3	0	2
JPN	2	1	1	2	0	1	87	5	0	0	1
KOR	1	5	2	0	1	0	9	81	1	0	0
SPA	2	0	2	0	1	8	2	1	78	1	5
TEL	0	1	0	0	18	1	2	1	1	73	3
TUR	4	0	2	2	0	2	2	4	4	0	80

Official test set confusion matrix with the full model. Accuracy on the test set is 81.5%

Accuracy

The full model that we used to classify the test set combines all features.

Feature Group	# Params	Accuracy(%)
POS	540,947	55.18
+ FreqChar	1,036,871	79.55
+ CharPrompt	2,111,175	79.82
+ Brown	5,664,461	81.09

Feature Group	# Params	Accuracy(%)
MAIN	5,664,461	81.09
MAIN + PsvRatio	5,664,472	81.00
MAIN + DocLen	5,664,472	81.09
MAIN + Punct	5,664,604	81.09
MAIN + Misspell	5,799,860	81.27
MAIN + CxtFxn	7,669,684	81.73
FULL MODEL	-	**84.55**

The accuracy for 10-fold cross-validation on the development set.

Analysis

Texts produced by non native English writers involves a tension between the imposing models of the native language, on the one hand, and a set of cognitive constraints resulting from the efforts to generate the target text, on the other. The former is called *interference* in Translation Studies. We explore the effects of interference by analysing several patterns we observe in the features.

- Arabic speakers use *a lot* as a single word more often and sometime omit the definite article before nouns and adjectives.
- German authors use Hyphens more frequently, probably due to compounding in their native language. They also tend to substitute the letter *y* with *z* and vice versa.
- Japanese authors confuse *l* and *r*.
- The characters *r* and *s* are misused in Chinese and Spanish, respectively.

Acknowledgements

This research was supported by a grant from the Israeli Ministry of Science and Technology.