

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. Our method uses standard text classification techniques based on multiclass logistic regression, combining individually weak indicators 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.

Features

For our classification model, we use the creg https://github.com/redpony/creg

Part-of-speech (POS) *n***-grams** Counts of every POS 1-, 2-, and 3-gram in each document.

FreqChar Counts of character *n*-grams that are observed more than 5 times in the training corpus. *n* ranges from 1 to 4.

CharPrompt Conjunction of the FreqChar features with the prompt ID

Brown clusters We clustered 8 billion words of English into 600 clusters of words that appear in similar left and right contexts. (http://www.ark.cs.cmu.edu/cdyer/en-600/cluster_viewer.html). Our features are counts of *n*-grams of Brown clusters as lexical representations. *n* ranges from 1 to 4.

PsvRatio The proportion of passive verbs out of all verbs.

DocLen Document length in tokens.

Punct 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. Highweight features include: ARA:DEL<e>, ARA:INS<e>; GER:SUBST<z>/<y>; JPN:SUBST<1>/<r>, JPN:SUBST<r>/<l>; SPA:INS<s>.

Restore Counts of substitutions, deletions and insertions of predefined tokens that we restored in the texts: CHI:DELWORD<do>;GER:DELWORD<on>.

CxtFxn Contextual function words. Counts of *n*-grams consisting of one or two function words and the POS tag of the adjacent words: CHI:<some JJ>; HIN:<as VBN>.

Example: L1 German sentence

Firstly the employers live more savely because they are going to have more money to spend for luxury.

	Presence	Considered alternatives	s/edits
Characters	"FreqChar_l_y_ ": "CharPrompt_P5_g_o_i": "Punct_period":	log 2 + 1 log 1 + 1 log 1 + 1 log 1 + 1 "Misspell_InsertP_p_,": "Misspell_MID:SUBST:v:f": "Misspell_SUBST:v:f":	1.0 1.0 $\log 1 + 1$ $\log 1 + 1$
	"DocLen_":	$log 19 + 1$ "Misspell_safely":	$\log 1 + 1$
	"MeanWordRank":	422.6 "Restore_Match_p_to":	0.5
Words	"CohMarker_because":	$log 1 + 1$ "Restore_Delete_p_to":	0.5
vvolus	"MostFreq_have":	$log 1 + 1$ "Restore_Delete_p_are":	1.0
	"PosToken_last_luxury":	$log 1 + 1$ "Restore_Delete_p_because":	1.0
	"Pronouns_they":	$log 1 + 1$ "Restore_Delete_p_for":	1.0
POS	"POS_VBP_VBG_TO":	$\log 1 + 1$	
103	"POS_p_VBP_VBG_TO":	0.059	
Words + POS	"CxtFxn_VBP_VBG_to":	$\log 1 + 1$	
	"CxtFxn_more_RB":	$\log 1 + 1$	
Brown	"C_1111101111110_110100011	110_110101100":	log 1 + 1

Brown clusters	Words in cluster
C_1111101111110	investors customers patients employees consumers users citizens share-
	holders clients individuals managers buyers viewers employers guests
	readers immigrants taxpayers humans donors households homeown-
	ers competitors travelers audiences borrowers shoppers offenders physi-
	cians creditors subscribers stockholders sellers entrepreneurs advertisers
	applicants motorists tenants builders smokers strangers collectors listen-
	ers savers retirees outsiders travellers bidders bondholders patrons
C_110100011110	live remain stay stand die sit compete operate invest participate arrive
	engage succeed lie cope gather testify comply communicate proceed
	weigh disagree cooperate intervene expire rein behave interact thrive in-
	terfere prevail persist coincide explode collaborate linger grips enroll in-
	dulge resonate dine tread prosper loom grapple reside retaliate collide
	regroup innovate
C_110101101	more less fewer

Results

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR	P(%)	R(%)
ARA	80	0	2	1	3	4	1	0	4	2	3	80.8	80.0
CHI	3	80	0	1	1	0	6	7	1	0	1	88.9	80.0
FRE	2	2	81	5	1	2	1	0	3	0	3	86.2	81.0
GER	1	1	1	93	0	0	0	1	1	0	2	87.7	93.0
HIN	2	0	0	1	77	1	0	1	5	9	4	74.8	77.0
ITA	2	0	3	1	1	87	1	0	3	0	2	82.1	87.0
JPN	2	1	1	2	0	1	87	5	0	0	1	78.4	87.0
KOR	1	5	2	0	1	0	9	81	1	0	0	80.2	81.0
SPA	2	0	2	0	1	8	2	1	78	1	5	77.2	78.0
TEL	0	1	0	0	18	1	2	1	1	73	3	85.9	73.0
TUR	4	0	2	2	0	2	2	4	4	0	80	76.9	80.0

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.

Main features	# 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

Additional features	# 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 + Restore	5,682,589	81.36
MAIN + CxtFxn	7,669,684	81.73
FULL MODEL	_	84.55

10-fold cross-validation on the development set.

Analysis

Texts produced by non native English writers involve 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 analyzing 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.