Identifying the L1 of Non-native Writers: The CMU-Haifa System

Yulia Tsvetkov[C], Naama Twitto[H], Nathan Schneider[C], Noam Ordan[H], Manaal Faruqui[C], Victor Chahuneau[C], Shuly Wintner[H], and Chris Dyer[C]
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA • 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.

Classification Model

We use a logistic regression classifier implemented by `creg` trained to maximize the log-likelihood of the training data, penalized by a combined ℓ_2 and entropic regularizer.

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \alpha \overbrace{\sum_j \lambda_j^2}^{\ell_2 \text{ reg.}} - \sum_{\{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{D}|}} \Bigg(\overbrace{\log p_{\boldsymbol{\lambda}}(y_i \mid \mathbf{x}_i)}^{\text{log likelihood}} + \tau \underbrace{\mathbb{E}_{p_{\boldsymbol{\lambda}}(y' \mid \mathbf{x}_i)} \log p_{\boldsymbol{\lambda}}(y' \mid \mathbf{x}_i)}_{-\text{conditional entropy}} \Bigg)$$

Features

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

FreqChar Counts of character 1–4-grams that are observed more than 5 times in the training corpus.

CharPrompt Conjunction of the FreqChar features with the prompt ID

Brown clusters We clustered 8 billion words of English into 600 clusters and used 1–4-grams.

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.

Restore Substitutions, deletions and insertions of common words that were restored with an n-gram LM.

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>`.

Example: L1 German sentence

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

Some of the features extracted:

		Presence		Considered alternatives/edits	
Characters	"FreqChar_l_y_ ":	$\log 2+1$	"Misspell_DeleteP_p_.":	1.0
	"CharPrompt_P5_g_o_i":	$\log 1+1$	"Misspell_InsertP_p_,":	1.0
	"Punct_period":	$\log 1+1$	"Misspell_MID:SUBST:v:f":	$\log 1+1$
			"Misspell_SUBST:v:f":	$\log 1+1$
Words	"DocLen_":	$\log 19+1$	"Misspell_safely":	$\log 1+1$
	"MeanWordRank":	422.6	"Restore_Match_p_to":	0.5
	"CohMarker_because":	$\log 1+1$	"Restore_Delete_p_to":	0.5
	"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":	$\log 1+1$		
	"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_110100011110_110101101100":		$\log 1+1$	

Brown clusters	Words in cluster
C_1111101111110	investors customers patients employees consumers users citizens shareholders clients individuals managers buyers viewers **employers** guests readers immigrants taxpayers humans donors households homeowners competitors travelers audiences borrowers shoppers offenders physicians creditors subscribers stockholders sellers entrepreneurs advertisers applicants motorists tenants builders smokers strangers collectors listeners 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 interfere prevail persist coincide explode collaborate linger grips enroll indulge resonate dine tread prosper loom grapple reside retaliate collide regroup innovate
C_110101101100	**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.