## Carnegie Mellon University

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

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer 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, native language of English test takers from the content of essays they write. Our method uses standard text classificati techniques based on multiclass logistic regression, combini individually weak indicators to predict the most probable tive language from a set of 11 possibilities: Arabic, Chine French, German, Hindi, Italian, Japanese, Korean, Spanish, 7 ugu, and Turkish.

### CLASSIFICATION MODEL

We use a logistic regression classifier implemented by cr trained to maximize the log-likelihood of the training data, nalized by a combined  $\ell_2$  and entropic regularizer.

$$\hat{\mathbf{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \alpha \sum_{j}^{\ell_{2}} \sum_{i}^{reg.} - \sum_{\{(\mathbf{x}_{i}, y_{i})\}_{i=1}^{|\mathcal{D}|}} \left( \underbrace{\log p_{\boldsymbol{\lambda}}(y_{i} \mid \mathbf{x}_{i})}_{\tau \mathbb{E} p_{\boldsymbol{\lambda}}(y' \mid \mathbf{x}_{i})} \log p_{\boldsymbol{\lambda}}(y' \mid \mathbf{x}_{i}) \right)$$

-conditional entropy

### FEATURES

Part-of-speech (POS) *n*-grams Counts of every POS 1-, 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 prompt ID
- **Brown clusters** We clustered 8 billion words of English i 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 substi tions, deletions, insertions and joinings.
- **Restore** Substitutions, deletions and insertions of comm words that were restored with an *n*-gram LM.
- **CxtFxn** Contextual function words. Counts of *n*-grams co sisting of one or two function words and the POS tag the adjacent words: CHI:<some JJ>.

		t	hey a	re goi	ing to	have	more	e mor	ney to	spen	d for	luxury	у.	
Som	e of t	he feat	tures ex	tracte	d:									
					I	Presence	e		(	Conside	ered alt	ernativ	es/edits	
	C	haracte	ers	"Cha	Char_l_y rPrompt_ ct_perioc	_P5_g_o_	_i": 1	og2+1 og1+1 og1+1	"N "N	/isspell_ /isspell_ /isspell_ /isspell_	InsertP_ MID:SU	_p_,": /BST:v:f":	1.0 1.0 : log1+1 log1+1	
	W	Vords		"Mea "Coh "Mos "PosT "Pron	Len_": nWordRa Marker_b tFreq_ha Token_las houns_th	because" ve": st_luxury ey":	2 : 1 1 v": 1 1	og19+1 422.6 og1+1 og1+1 og1+1 og1+1	"N "F "R "R "R	/isspell_ lestore_N lestore_I lestore_I	safely": Match_p Delete_p Delete_p Delete_p	o_to": o_to": o_are": o_becaus	log1+1 0.5 0.5 1.0	
	Р	OS			_VBP_VE _p_VBP_			og 1 + 1 ).059						
	W	Vords +	POS	"CxtF	 Fxn_VBP_ Fxn_more	_VBG_to'	": 1	og1+1 og1+1						
	B	rown			1101111		.000111	10_1101	0110110	0":		log1	+1	
B	roum	cluster						Mor	ds in clu	letor				
		$\frac{\text{cluster}}{1011111}$					. •					• . •	ens shar	1 1 1
C	_1101(	0001111	bo en co	orrowers trepren llectors	s shopp leurs ad listene	ers offe lvertiser rs savers	nders s appl s retire	physici licants es outsi	ans cre motoris iders tra	ditors s sts tena wellers	subscril nts bui bidder	oers sto ilders sr s bondh	velers au ckholder nokers s olders pa rrive eng	trange atrons
		0001111	bo en co l0 liv ce int co ret	orrowers trepren llectors <b>e</b> rema ed lie c tervene llabora taliate c	s shopp leurs ad listener in stay ope gat expire	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i	enders is appl s retire ie sit c ify con nave in nroll in	physici licants es outsi ompete nply con teract th idulge re	ans cre motoris iders tra operat mmunie hrive in	ditors s ts tena vellers e inves cate pro terfere	subscril nts bui bidders t partic oceed w prevail	oers sto ilders sr s bondh cipate ar veigh di persist	ckholden mokers s iolders pa	cs sell strang atrons age su copera explo
			bo en co l0 liv ce int co ret	orrowers trepren llectors <b>e</b> rema ed lie c tervene llabora taliate c	s shopp leurs ad listener in stay ope gat expire te linger collide re	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i	enders is appl s retire ie sit c ify con ave in nroll in nnova	physici licants es outsi ompete nply con teract th idulge re	ans cre motoris iders tra operat mmunie hrive in	ditors s ts tena vellers e inves cate pro terfere	subscril nts bui bidders t partic oceed w prevail	oers sto ilders sr s bondh cipate ar veigh di persist	ckholden nokers s olders pa rrive eng sagree co coincide	cs sell strang atrons age su copera explo
C			bo en co l0 liv ce int co ret	orrowers trepren llectors <b>e</b> rema ed lie c tervene llabora taliate c	s shopp leurs ad listener in stay ope gat expire te linger collide re	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i	enders is appl s retire ie sit c ify con ave in nroll in nnova	physici licants es outsi ompete nply con teract th dulge re te	ans cre motoris iders tra operat mmunie hrive in	ditors s ts tena vellers e inves cate pro terfere	subscril nts bui bidders t partic oceed w prevail	oers sto ilders sr s bondh cipate ar veigh di persist	ckholden nokers s olders pa rrive eng sagree co coincide	cs sell strang atrons age su opera ole resi
C_ tr	- _1101( rue↓ ARA	0110110	bo en co l0 liv ce int co ret	orrowers trepren llectors re rema ed lie c tervene llabora taliate c ore less	s shopp leurs ad listener in stay s ope gat expire r te linger collide re	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i	enders s appl s retire ie sit c ify con ave in nroll in nnova	physici licants es outsi ompete nply con teract th dulge re te ULTS JPN 1	ans cre motoris iders tra coperate mmunie hrive in esonate	ditors s sts tena vellers e inves cate pro terfere dine tre	subscril nts bui bidders t partic oceed w prevail ead pro	persist osper loc	ckholden nokers s lolders pa rrive eng sagree co coincide om grapp P (%) 80.8	s selle strange atrons age su opera ole resi ole resi <b>R (9</b> 80.
-C_ tr f	- _1101( rue↓ ARA CHI	0110110 ARA 80 3	bo en co co l0 liv ce int co ret 00 mo	orrowers trepren llectors e rema ed lie c tervene llabora taliate c ore less FRE 2 0	s shopp leurs ad listener in stay s ope gat expire r te linger collide re fewer fewer	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i	enders s appl s retire ie sit c ify con ave in nova nova innova ITA 4 0	physici licants es outsi ompete nply con teract th dulge re te	ans cre motoris iders tra coperate mmunie hrive in esonate	ditors s sts tena vellers e inves cate pro terfere dine tr dine tr SPA 4 1	subscril nts bui bidders t partic oceed v prevail ead pro	oers sto ilders sr s bondh cipate an veigh di persist sper loc TUR 3 1	ckholder mokers s olders pa rive eng sagree co coincide om grapp P (%) 80.8 88.9	s selle strange atrons age su opera explo ole resi R (9 80. 80.
C_ tr f (	- 11010 rue↓ ARA CHI FRE	0110110 ARA 80	bo en co l0 liv ce int co ret 00 mo	orrowers trepren llectors re rema ed lie c tervene llabora taliate c ore less	s shopp leurs ad listener in stay s ope gat expire f te linger collide re fewer fewer fewer	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i	enders is appl s retire ie sit c ify con ave in nroll in nnova RES ITA 4	physici licants es outsi ompete nply con teract th dulge re te ULTS JPN 1	ans cre motoris iders tra operat mmunie hrive in esonate	ditors s sts tena vellers e inves cate pro terfere dine tre SPA	subscril nts bui bidders t partic oceed w prevail ead pro TEL 2 0 0	oers sto ilders sr s bondh cipate an veigh di persist osper loc TUR 3 1 3	ckholder mokers s olders pa rive eng sagree co coincide om grapp 80.8 88.9 88.9 86.2	s sell strang atrons age si oper explo ole res R (9 80. 80. 80. 81.
C_ C_ fr f ( ]	- 11010 rue↓ ARA CHI FRE GER	O110110 ARA 80 3 2 1	bo en co co l0 liv ce int co ret 00 mo	orrowers trepren llectors e rema ed lie c tervene llabora taliate c ore less FRE 2 0	s shopp leurs ad listener in stay s ope gat expire r te linger collide re fewer fewer	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i HIN 3 1 1 1 0	enders s appl s retire ie sit c ify con ave in nova nova innova ITA 4 0	physici licants es outsi ompete nply con teract th dulge re te ULTS JPN 1	ans cre motoris iders tra operat mmunie hrive in esonate	ditors s sts tena vellers e inves cate pro terfere dine tr dine tr sPA 4 1 3 1	subscril nts bui bidders t partic oceed v prevail ead pro TEL 2 0 0 0	oers sto ilders sr s bondh cipate an veigh di persist sper loc TUR 3 1	ckholder nokers s olders pa rrive eng sagree co coincide om grapp 80.8 88.9 88.9 86.2 87.7	s sell strang atrons age si oper explo ole res 80. 80. 80. 81. 93.
C_ C_ fr f ( ]	- 11010 rue↓ ARA CHI FRE GER HIN	0110110 ARA 80 3	bo en co co int ce int co ret 00 m e	orrowers trepren llectors e rema ed lie c tervene llabora taliate c ore less FRE 2 0 81 1 0	s shopp leurs ad listener in stay s ope gat expire f te linger collide re fewer fewer fewer	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i	enders s appl s retire ie sit c ify con ave in nroll in nova ITA 4 0 2 0 2 1	physici licants es outsi ompete nply con teract th dulge re te ULTS JPN 1	ans cre motoris iders tra operat mmunie hrive in esonate	ditors s sts tena vellers e inves cate pro terfere dine tr dine tr dine tr 1 3 1 1 5	subscril nts bui bidders t partic oceed v prevail ead pro TEL 2 0 0 0 0 9	oers sto ilders sr s bondh cipate an veigh di persist osper loc TUR 3 1 3	ckholden nokers s olders pa rrive eng sagree co coincide om grapp 80.8 88.9 88.9 88.9 88.2 87.7 74.8	s sell strang atrons age s oper explo ole res 80. 80. 80. 80. 81. 93. 77.
C_ C_ E	- 11010 rue↓ ARA CHI FRE GER HIN ITA	D110110 ARA 80 3 2 1 2	bo en co co l0 liv ce int co ret 00 mo	orrowers trepren llectors e rema ed lie c tervene llabora taliate c ore less FRE 2 0	s shopp leurs ad listener in stay s ope gat expire r te linger collide re fewer fewer fewer	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i HIN 3 1 1 1 0	enders s appl s retire ie sit c ify con ave in nova nova innova ITA 4 0	physici licants es outsi ompete nply con teract th dulge re te ULTS JPN 1 6 1 1 6 1 1 0	ans cre motoris iders tra operat munie hrive in esonate KOR 0 7 0 1 1 0	ditors s sts tena vellers e inves cate pro terfere dine tr dine tr 5 3 1	subscril nts bui bidders t partic oceed v prevail ead pro TEL 2 0 0 0 0 0 0	Sers sto ilders sr s bondh cipate an veigh di persist sper loc sper loc 1 3 1 3 2 4	ckholder nokers s olders pa rrive eng sagree co coincide om grapp 80.8 88.9 88.9 88.9 88.2 87.7 74.8 82.1	s sell strang atrons age s oper explo ole res 80. 80. 80. 80. 81. 93. 77. 87.
	I 11010 I 1100 I 1100 I 1100 I 1100 I 1100 I 1100 I 1100 I 1100 I	O110110 ARA 80 3 2 1 2 1 2 2	bo en co co l0 liv ce int co ret 0 m 2 1 0 80 2 1 1 0 0 1	orrowers trepren llectors re rema ed lie c tervene llabora taliate c ore less ore less fRE 2 0 81 1 1 0 3 1	s shopp leurs ad listener in stay s ope gat expire r te linger collide re fewer fewer fewer gas 1 1 1 1 2	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i HIN 3 1 1 1 0	enders s appl s retire ie sit c ify con ave in nroll in nova ITA 4 0 2 0 1 87 1	physici licants : es outsi ompete nply con teract t dulge r te te ULTS JPN 1 6 1 6 1 6 1 6 1 1 6 1 1 87	ans cre motoris iders tra operat munie hrive in esonate KOR 0 7 0 1 1 0 5	ditors s sts tena vellers e inves cate pro terfere dine tr dine tr dine tr 1 3 1 1 5	subscril nts bui bidders t partic oceed v prevail ead pro TEL 2 0 0 0 0 0 0 0 0	Sers sto ilders sr s bondh cipate an veigh di persist sper loc sper loc 1 3 1 3 2 4	Ckholden mokers s lolders pa rrive eng sagree co coincide om grapp P (%) 80.8 88.9 86.2 87.7 74.8 82.1 78.4	s sell strang atrons age si oper explo ole res 80. 80. 80. 81. 93. 77. 87. 87. 87.
C_ C_ tr f C_ I C_ I K	- - - - - - - - - - - - - -	O110110 ARA 80 3 2 1 2 1 2 2	bo en co co int ce int co ret 00 m e	orrowers trepren llectors e rema ed lie c tervene llabora taliate c ore less FRE 2 0 81 1 0	s shopp eurs ad listener in stay s ope gat expire r collide re fewer fewer fewer gas fa f	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i HIN 3 1 1 1 0	enders s appl s retire ie sit c ify con ave in nroll in ave in nova ITA 4 0 2 0 1 87 1 87 1 0	physici licants es outsi ompete nply con teract th dulge re te ULTS JPN 1 6 1 1 6 1 1 0	ans cre motoris iders tra operat munie hrive in esonate KOR 0 7 0 1 1 0	ditors s sts tena vellers e inves cate pro terfere dine tr dine tr 5 3 1 3 1 1 3 1 1	subscril nts bui bidders t partic oceed v prevail ead pro TEL 2 0 0 0 0 0 0	Sers sto ilders sr s bondh cipate an veigh di persist sper loc sper loc sper loc 1 1 3 2 1 1 3 2 1 1 3 1 1 3 1 1 3 1 1 3 1 1 3 1 1 3 1 1 3 1 1 3 1 1 1 3 1 1 1 3 1 1 1 1 1 1 0	ckholder mokers s olders pa rrive eng sagree co coincide om grapp 80.8 88.9 88.9 88.9 88.2 87.7 74.8 82.1 78.4 80.2	s sell strang atrons age s oper explo le res 80. 80. 80. 81. 93. 77. 87. 87. 87. 81.
	I 11010 I 1100 I 1100 I 1100 I 1100 I 1100 I 1100 I 1100 I 1100 I	D110110 ARA 80 3 2 1 2 1 2 2 1 2 1	bo en co co l0 liv ce int co ret 0 m 2 1 0 80 2 1 1 0 0 1	FRE   1   0   1   1   1   1   1   1   2   1   2   1   2   1   2   1   2   1   2   1   2   1   2	s shopp leurs ad listener in stay s ope gat expire r te linger collide re fewer fewer fewer gas 1 1 1 1 2	ers offe lvertiser s savers stand di her testi rein beh grips er egroup i HIN 3 1 1 1 0	enders s appl s retire ie sit c ify con ave in nroll in nova ITA 4 0 2 0 1 87 1	physici icants : es outsi ompete nply con teract t dulge r te te ULTS JPN 1 6 1 6 1 1 6 1 1 6 1 1 6	ans cre motoris iders tra operat munie hrive in esonate KOR 0 7 0 1 1 0 5	ditors s sts tena vellers e inves cate pro terfere dine tr dine tr 5 3 1	subscril nts bui bidders t partic oceed v prevail ead pro TEL 2 0 0 0 0 0 0 0 0	Sers sto ilders sr s bondh cipate an veigh di persist sper loc sper loc 1 3 1 3 2 4	Ckholden mokers s lolders pa rrive eng sagree co coincide om grapp P (%) 80.8 88.9 86.2 87.7 74.8 82.1 78.4	s sell strang atrons age su opera ole resi ole resi R (9 80.

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

### EXAMPLE: L1 GERMAN SENTENCE

Main features	# Params	Accuracy (%)
POS only	540,947	55.18
+ FreqChar	1,036,871	79.55
+ CharPrompt	2,111,175	79.82
+ Brown	5,664,461	81.09
Brown only	1,240,420	72.26
		·
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	8,312,205	84.55

Main features	# Params	Accuracy (%)
POS only	540,947	55.18
+ FreqChar	1,036,871	79.55
+ CharPrompt	2,111,175	79.82
+ Brown	5,664,461	81.09
Brown only	1,240,420	72.26
		·,
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	8,312,205	84.55

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.





### ACCURACY

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

10-fold cross-validation on the development set.

### ANALYSIS

• Arabic speakers use *alot* 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.