Supplementary Material
Estimating Code-Switching on Twitter with a Novel Generalized Word-Level Language Detection Technique

Shruti Rijhwani
Language Technologies Institute
Carnegie Mellon University
`srijhwan@cs.cmu.edu`

Royal Sequiera
University of Waterloo
`rdsequie@uwaterloo.ca`

Monojit Choudhury **Kalika Bali** **Chandra Sekhar Maddila**
Microsoft Research India
{`monojitc,kalikab,chmaddil`}`@microsoft.com`

1 Error Analysis

We conducted a thorough analysis of the errors from all languages. *nl* words marked as *en* account for nearly 14% of all errors. We observe that most of these are actually *en* words with *nl* gold-standard labels, which is the convention used by dataset creators. GWLD labels these *en* words accurately. We also observe that 13% and 7% of the word-level errors come from confusion between *es-en* and *nl-tr* respectively. A large number of these are named entities (`Twitter`, `Orhan Pamuk`) and ambiguous words (`a`, `no`). 41% of the *es-en* errors are undetected single words language switches. This is because GWLD is inclined to remain in the same language for unseen words. It must be noted that the GWLD accurately labels over 70% of all single-word code-switching in es-en, including ambiguous and misspelled instances. Confusion between *pt* and *es* contribute 10% of the total errors because these languages have several common words. The language pairs not discussed account for less than 4% of the errors each.

GWLD sometimes detects languages that are not present in the tweet, which account for a sizable fraction (39.6%) of all word-level errors. Several of these overlap with the errors discussed previously and the causes are similar – named entities, misspelled words and ambiguous words. Single-alphabet tokens, which often belong to more than one language (`a`, `y`) or may be meaningless (`g`, `z`), cause 5.9% of all errors.

GWLD not detecting a language switch causes 7.7% of the word-level errors. 93.5% of these errors occur with fragments containing less than 3 words. As noted earlier, the GWLD generally performs well for such short phrases and the mislabeled instances typically contain out-of-vocabulary and ambiguous words. In fact, this is a desirable property of the system because, if an unseen word is encountered within a fragment of language l_i, it is better to label it as l_i rather than hypothesizing it to be from some $l_j \neq l_i$.

Language	Fraction
en	.741
es	.095
fr	.037
pt	.035
tr	0.031
de	0.016
nl	0.009
code-switched	0.036

Table 1: Worldwide Language Distribution

Language Pair	Fraction
en-es	.215
en-fr	.208
en-pt	.183
en-nl	.096
en-de	.09
en-tr	0.061
es-fr	0.032
fr-pt	0.012
other	.103

Table 2: Worldwide CS Distribution

2 Code-switching Statistics

This section provides more detailed statistics on the distribution of tweets in the corpus we use to analyze code-switching patterns.

City	Tweets	TopMonolingual	MixAmt	TopMixed
San Francisco	532K	en .94, es .02	.02	en-es .26, en-fr .19
New York City	690K	en .94, es .02	.02	en-es .21, en-es .19
San Diego	432K	en .86, es .09	.02	en-es .29, en-nl .14
Miami	290K	en .9, es .04	.02	en-es .33, en-pt .20
Houston	588K	en .96, es .01	.01	en-es .22, en-fr .21
Toronto	136K	en .94, pt .01	.02	en-fr .29, en-pt .19
Montréal	26K	en .68, fr .22	.05	en-fr .41, es-fr .19
Québec City	110K	fr .54, en .34	.08	en-fr .47, es-fr .22
Mexico City	332K	es .79, en .10	.07	en-es .54, es-fr .14,
Rio de Janeiro	1.7M	pt .90, en .03	.04	en-pt .52, fr-pt .16
Buenos Aires	470K	es .89, en .04	.03	en-es .43, es-fr .29
London	492K	en .94, es .01	.02	en-fr .26, en-pt .17
Paris	158K	fr .7, en .18	.07	en-fr .43, es-fr .21
Frankfurt	74K	de .52, en .29	.06	en-de .54, en-tr .07
Leipzig	4.3K	de .68, en .21	.05	en-de .64, de-tr .07
Berlin	23K	en .52, de .32	.06	en-de .53, de-tr .05
Amsterdam	310K	en .47, nl .40	.03	en-nl .41, en-pt .08
Lisbon	476K	pt .73, en .14	.06	en-pt .50, fr-pt .14
Geneva	11K	en .55, fr .29	.04	en-fr .46, es-fr .11
Zürich	9K	en .53, de .29	.05	en-de .45, en-fr .18
Brussels	100K	fr .44, en .42	.06	en-fr .37, es-fr .15
Madrid	147K	es .83, en .08	.06	en-es .43, es-fr .32
Barcelona	85K	es .71, en .19	.05	en-es .53, es-fr .17
Istanbul	351K	tr .61, en .21	.12	en-tr .53, nl-tr .13

Table 3: Tweet Language Distribution over Cities. Col: Tweets - total number of tweets analyzed; TopMonolingual - the top two languages with largest amount of monolingual tweets along with their fraction of the total tweets from the city; MixAmt - fraction of CS tweets from the city; TopMixed - the top two most frequently code-switched pairs along with the fraction of the CS tweets in these pairs among all the CS tweets from the city.

- Table 3 shows the city-wise top 2 monolingual and top 2 CS languages in terms of tweet fractions. All 24 cities that we obtained tweets from are reported here.
- Table 1 shows the language distribution in the corpus of tweets collected globally. 3.6% of those are code-switched.
- Table 2 shows the distribution within the code-switched tweets in the global tweet corpus. The top 8 mixed languages are shown and the other languages count for less than 1% of the total CS tweets.