Developing Language-tagged Corpora for Code-switching Tweets

Suraj Maharjan[1], Elizabeth Blair[2], Steven Bethard[2], and Thamar Solorio[1]

[1]Department of Computer Science
University of Houston
Houston, TX, 77004
smahajan2@uh.edu solorio@cs.uh.edu

[2]Department of Computer and Information Sciences
University of Alabama at Birmingham
Birmingham, Alabama 35294-1170
{eablair, bethard}@uab.edu

Introduction

- Code-switching :
  - Switching between languages within single context
  - Inter and intra sentential code-switching

No! No! No!

@user tyo school*ko* nam k ho?

Ronaldo le ta ramro game khelyo hai.

@user ammo bichara ma :(sab dos chai malai, very sad :'(:P

@user @user no Graciela Eso me Lo Dicen everyday so there's no need

Por favor **Gabriela** stop this right now!!!!

@user haha. info ta collect gareeko *raichas ta* :P

lang1
lang2
ambiguous
NE
mixed
other

Applications

- Law Enforcement : Mexico-USA drug trafficking
- Child Language : Analysis in multi-lingual environments
- Data mining : Ignoring multilingual may result in significant lost in business opportunities

Data Distribution

Language Pair	Training	Test	User Information
Nepali-English	9,993	2,874	15 males and 6 females from Kathmandu
Spanish-English	11,400	3,014	9 males and 11 females from Eastern, Central, Pacific, Mountain (US & Canada) time zones

Table 1: Tweets distribution for training and test set.

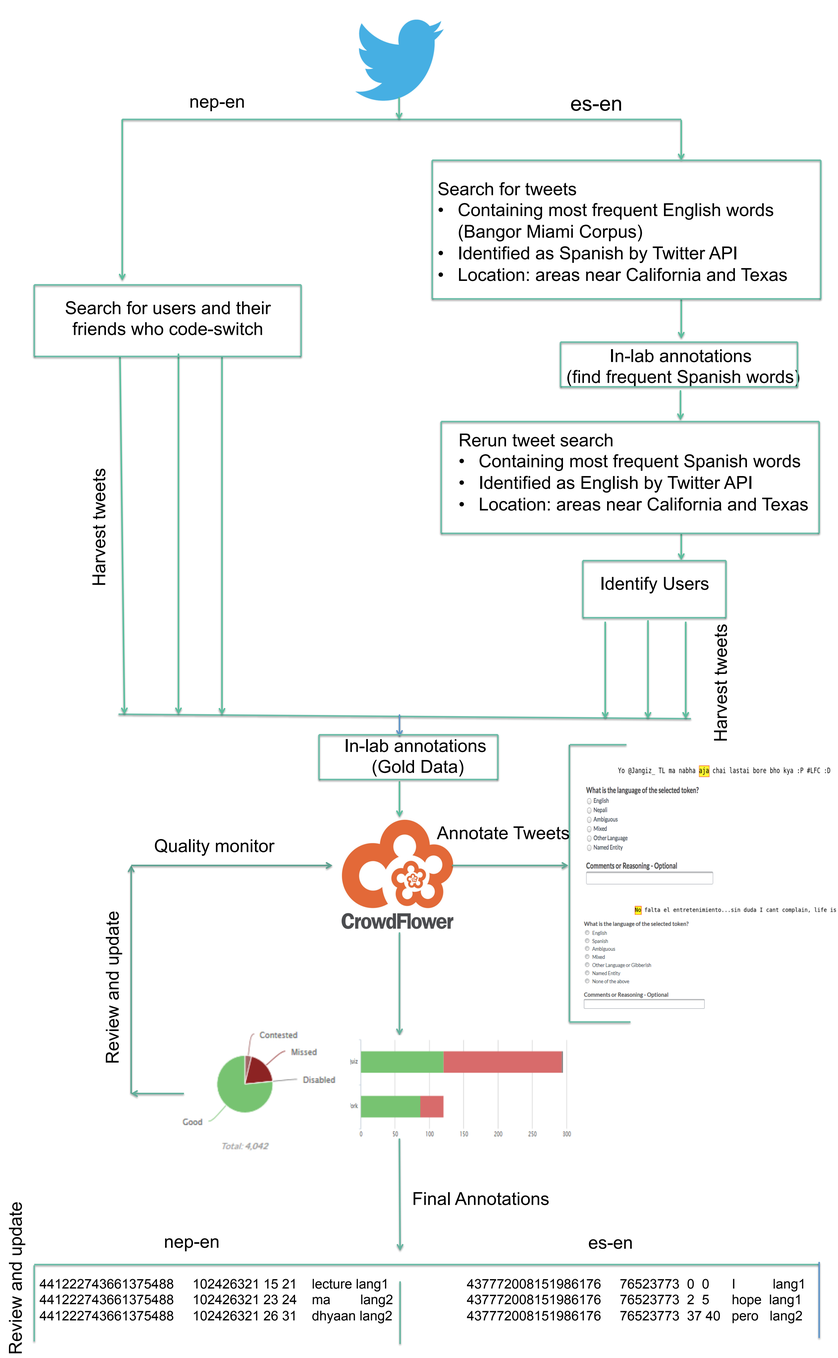
Tag	Nepali-English		Spanish-English	
	Training (%)	Test (%)	Training(%)	Test (%)
Lang1	31.14	19.76	54.78	43.28
Lang2	41.56	49.1	23.52	30.34
Mixed	0.08	0.60	0.04	0.03
NE	2.73	4.19	2.07	2.22
Ambiguous	0.09	-	0.24	0.12
Other	24.41	26.35	19.34	24.02

Table 2: Distribution of tags across training and test datasets.

PDF vs Inline Instructions for Annotators

- 1,000 tweets using PDF instruction scheme and 500 tweets using inline instruction scheme were reviewed
- Inter-annotators agreement
  - At most 0.01 annotators agreement difference between PDF and inline instruction schemes
  - Fleiss multi-Π, Cohen multi-κ

Workflow



Results

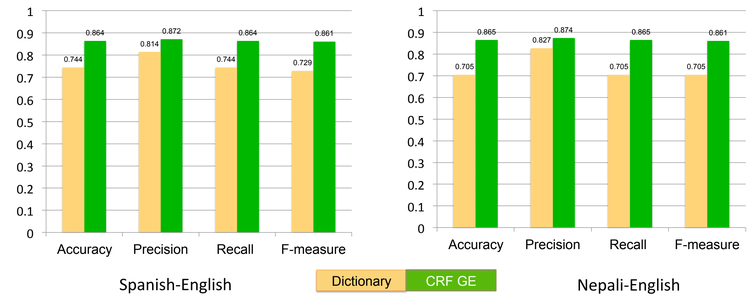


Fig 1: Benchmark system performance at the word level.

- Dictionary
  - Build separate dictionaries for each of the label from training dataset
  - Use dictionaries to assign label for each token
  - Use majority language to break the ties
- CRF GE
  - Train LangID (King and Abney) with monolingual texts for each language
  - English-Spanish training set : Tweets pulled from Texas and California
  - Nepali monolingual training set : Nepali song lyrics, news website and Romanized Nepali tweets

Analysis

Tokens	Nepali-English(%)	Spanish-English(%)
words	1.39	3.54
char-2	52.01	52.21
char-3	33.36	40.36
char-4	12.66	21.31
char-5	3.43	9.00

Table 3: N-gram overlap across language pairs.

- Dictionary based approach had problems with
  - Similar spelling : man (like, heart), gate (date), din (day), me, red
  - Unseen tokens : b-lated, yuss, comrade
  - No standard Romanized spelling for Nepali words
- CRF GE failed to detect
  - Small code-switched content embedded inside large monolingual segments
- Both bilinguals frequently used English function words (the, to, yo, he, she, and) and abbreviations used in social media (lol, lmao, idk)

Conclusion

- Code Switching is prevalent, complex and growing
- Corpus has not only been tagged for languages but also for named entities, mixed, ambiguous and irrelevant characters
- Our methods for searching and locating code-switching users can be helpful

Partially supported by

We thank Mona Diab and Julia Hirschberg for their contributions to the general annotation guidelines as well as the in lab annotators and CrowdFlower contributors for their labeling efforts.

LAW IX,
June 05, 2015,
Denver, Colorado