Toward Web-scale Analysis of Codeswitching

Constantine Lignos and Mitch Marcus (University of Pennsylvania, Johns Hopkins University)

human language technology
center of excellence

Background

Goal: Create tools to enable large-scale analysis of codeswitching

Codeswitching (CS)

- *Concurrent use of more than one language*
- Differentiated from *borrowing*:
  (1) i'm at work babe. trabajando *like a good girl*. que haces. estoy entusiamada porque voy al *mall* con gina!!!
  "I'm at work babe. Working like a good girl. What are you doing? I'm excited because I'm going to the mall with Gina!"
- Sociolinguistic significance and known formal constraints (e.g., Belazi et al. 1994, Poplack 1980)

Using social media

- Current spontaneous CS data sets are small corpora collected through sociolinguistic interviews
- Lots of social media data, many challenges to finding CS:
  - Non-standard spelling, capitalization, punctuation
  - Heavy borrowing of transient forms: brand names, technology terms, movie and song titles
  - *No annotated data*

Building a corpus of CS

- Used separate primarily Spanish (6.8 million) and English (2.8 million) tweets to model each language
- Tested using Spanish corpus collected by MITRE: 93,136 tweets (Burger et al. 2011)
- Developed *Codeswitchador* to identify CS, labeled about 11% as CS (10,020)
- Accuracy verified using test set created by bilingual Mechanical Turk annotators
  - Instructed to account for standard borrowings
  - Token inter-annotator agreement: 96.6%

Codeswitchador for labeling CS

Ratio list model

- Using two (mostly) monolingual corpora, ratio of probability of a word *w* in each:

$$\frac{p(w \mid Spanish)}{p(w \mid English)}$$

- Label each word by dominant language, but if ratio is near 1.0 treat as ambiguous
- If enough words from each language, mark tweet as CS

Ambiguous/unknown words

- Ambiguous and out-of-vocabulary (OOV) words must be disambiguated by context
- Best-performing approach: match word to the left
- Why? *Functional Head Constraint* (Belazi et al. 1994)
  - Functional heads match their complement

Performance

- Correct language for each word? 96.9% (baseline 92.3%)
- Is a tweet codeswitched? Prec.: .951, Recall: .922, F1: .936

Why so simple?

- Fast, can give per-item confidence
- More complex models (HMM, SVM-HMM) do not lead to significantly better performance, require annotated data
- Future work: integrate morphological information and part-of-speech tags

Most common errors

- English words as brand names: Apple, Blackberry
- Standard cognates/borrowings: radio, piano, marketing

Use it:

https://github.com/ConstantineLignos/Codeswitchador

Analysis of tweets

Support for the func. head constraint:

Complementizer *always* matches its complement:

- All cases of apparent mismatch are proper names (Apple):
  (2) Yo no creo que *apple* sace un Tablet Mac.
  "I don't think Apple will release a Tablet Mac."

Determiner *usually* matches its complement:

- Most exceptions are idioms or lexical gaps:
  (3) la *memory stick* "the memory stick"
  (4) la *big sister* del tuicter "the big sister of Twitter"

- True exceptions:
  (5) hoy fue el *baby shower* de mi *wife*
  "Today I went to my wife's baby shower"
  [*mi* and *wife* predicted to be the same language]

(6) *here* siento el *fragrance* of las *flower* el *sound* of las gabiotas las olas del *sea* ... estoy en el *paradise*
"Here I feel the fragrance of the flowers, the sound of the seagulls, the waves of the sea…I'm in paradise"

Extensions:

- Use corpora to create grammars of CS (cf. Sankoff and Poplack 1981)
- Automated extraction of frequencies for typology of codeswitching: sentential, phrasal, sub-phrasal
- Correlate usage of CS with sociolinguistic factors

References

Belazi, Hedi M., Edward J. Rubin and Almeida Jacqueline Toribio (1994). Code Switching and X-Bar Theory: The Functional Head Constraint. *Linguistic Inquiry* 25:2, 221-237.

Burger, John D., John Henderson, George Kim, and Guido Zarrella (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1301-1309.

Poplack, Shana. (1980). Sometimes I'll start a sentence in Spanish y termino en español. *Linguistics* 18: 581-618.

Sankoff, David and Shana Poplack (1981). A formal grammar for code-switching. *Research on Language & Social Interaction*, 14: 1, 3-45.