INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

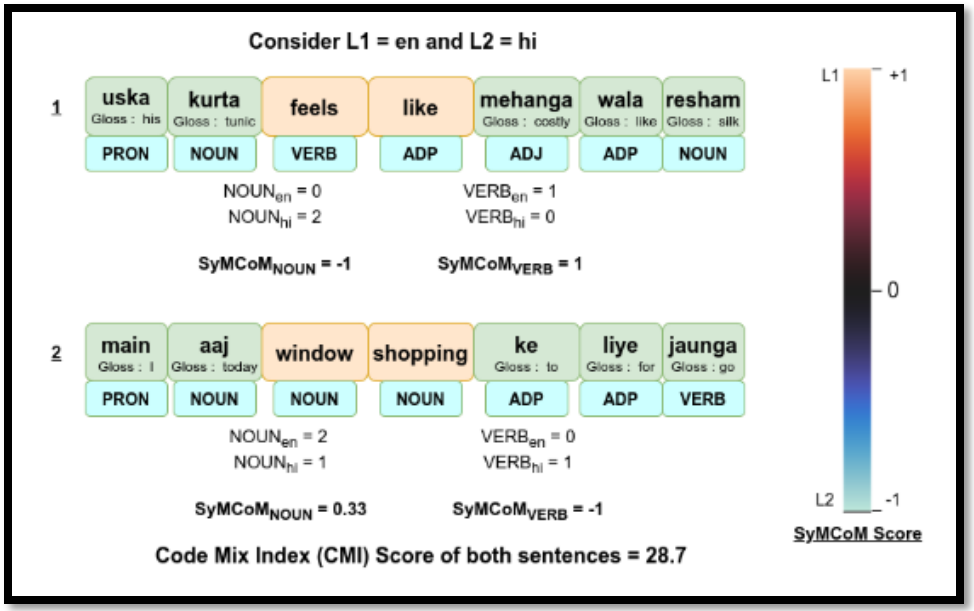
LTRC, IIIT Hyderabad

SyMCoM - Syntactic Measure of Code Mixing

A Study Of English-Hindi Code-Mixing

Prashant Kodali[†] **Anmol Goel**[†] **Monojit Choudhury**[‡]
Manish Shrivastava[†] **Ponnurangam Kumaraguru**[†]
[†]International Institute of Information Technology Hyderabad
[‡]Microsoft Research, India

Introduction and Motivation



- Code mixing - bilingual speakers tend to switch between two or more languages in conversations.
- For measuring variety of code mixing in, and across corpus, Language ID (LID) tags based measures (CMI) have been proposed.
- Syntactical variety/patterns of code-mixing and their relationship vis-a-vis computational model's performance is under explored

SyMCoM: Syntactic Measure of Code Mixing

- We propose **SyMCoM**, a metric to measure syntactic variety in code-mixed text.

- SyMCoM can be computed at multiple levels
  - for a syntactical unit (PoS, chunks, phrases, clauses)
  - for a sentence
  - for a corpus

$$SyMCoM_{SU} = \frac{(Count_{SU_{L1}}) - (Count_{SU_{L2}})}{\sum_{i=1}^{2} Count_{SU_{Li}}} \quad (1)$$

$$SyMCoM_{sent} = \sum_{SU} \frac{Count_{SU}}{len} \times |SyMCoM_{SU}| \quad (2)$$

$$SyMCoM_{corpus} = \sum_{sent} \frac{SyMCoM_{sent}}{\#\ sentences\ in\ corpus} \quad (3)$$

SyMCoM can differentiate between sentences with similar CMI scores

Ex 1
$SyMCoM_{NOUN,ADJ}$ = -0.76
$SyMCoM_{VERB,ADV}$ = 0.46
CMI = 40

dimaag NOUN | ka ADP | baaja NOUN | baja VERB | before SCONJ | i PRON | realized VERB | you PRON | were AUX | kkidding VERB

Ex 2
$SyMCoM_{NOUN,ADJ}$ = 0.99
$SyMCoM_{VERB,ADV}$ = -0.46
CMI = 40

last ADJ | day NOUN | pe ADP | first ADJ | day NOUN | wale ADP | posts NOUN | like NOUN | kar VERB | dena AUX

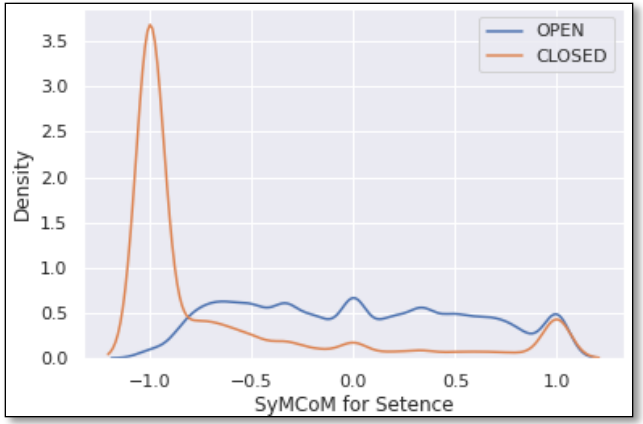
Ex 3
$SyMCoM_{NOUN,ADJ}$ = 0
$SyMCoM_{VERB,ADV}$ = -0.76
CMI = 28.5

ali PROPN | azmat PROPN | ki ADP | awaaz NOUN | will AUX | always ADV | give VERB | me PRON | goosebumps NOUN

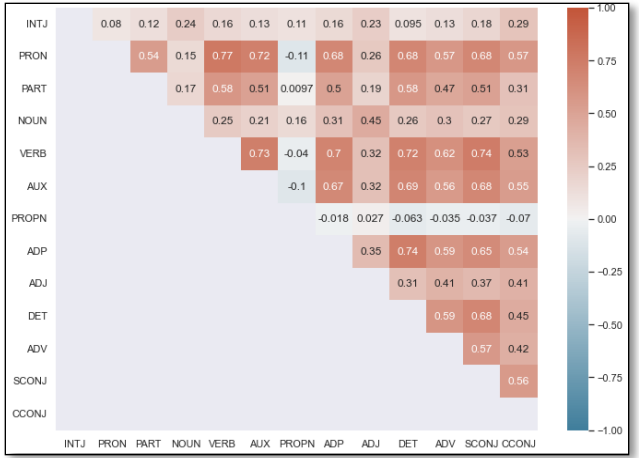
Ex 4
$SyMCoM_{NOUN,ADJ}$ = 0.76
$SyMCoM_{VERB,ADV}$ = -0.76
CMI = 12.5

this DET | chamcha NOUN | akways ADV | has VERB | a DET | nontreatable ADJ | verbal ADJ | diarrhoea NOUN

Observations

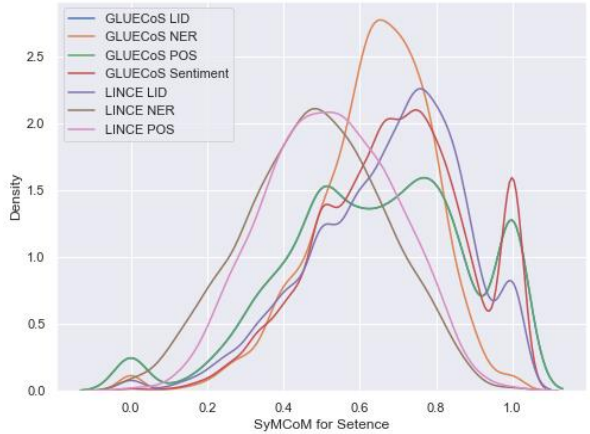


Open class categories (e.g., noun, adjectives) are more likely to be switched than the closed class categories (e.g., pronouns, verbs) within a sentence.

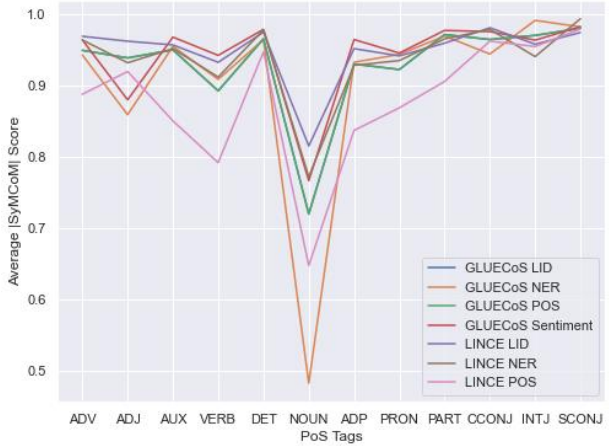


SyMCoM scores for verb are highly correlated with closed class words.

Observations



The plot represents the syntactic variation across benchmark datasets which encode the switching within PoS tag categories.



The plot represents mixing specific to each PoS tag. Across benchmarks, NOUN is highly switched, followed by VERB. But other PoS tags are largely monolingual.

Limitations and Future Directions

- Extending SyMCoM to
  - code-mixing between 3 or more languages and
  - to deeper syntactic structures (nested phrases) are left as part of future work.
  - Further analysis of time ordering of syntactic patterns could lead to fruitful observations.

- Establishing relationship between SyMCoM (and other Code Mix metrics) and downstream task performance

- Limitation: Validity of SyMCoM scores largely dependent on PoS tagger. Extension/application of SyMCoM to other language pairs are limited on availability of such resource.