SyMCoM - Syntactic Measure of Code Mixing

A Study Of English-Hindi Code-Mixing

Prashant Kodali[1], Anmol Goel[1], Monojit Choudhury[2], Manish Shrivastava[1], Ponnurangam Kumaraguru[1]
[1] IIIT Hyderabad, [2] Microsoft Research

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD
LTRC, IIIT Hyderabad

IIIT HYDERABAD

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 variety/patterns of code-mixing and their relationship vis-a-vis computational model's performance is under explored

Contributions

- In this work, we have proposed SyMCoM , a syntax-aware measure of code-mixing, to analyze code-mixed corpora from a syntactic perspective.

- Our analysis confirms a few important tenets of the matrix language theory, including the fact that CLOSED class categories and (finite) verbs are less likely to be switched.

- Additionally, we have trained a English-Hindi (Hinglish) PoS Tagger using XLM-R which is able to achieve state-of-the-art-results.

SyMCoM

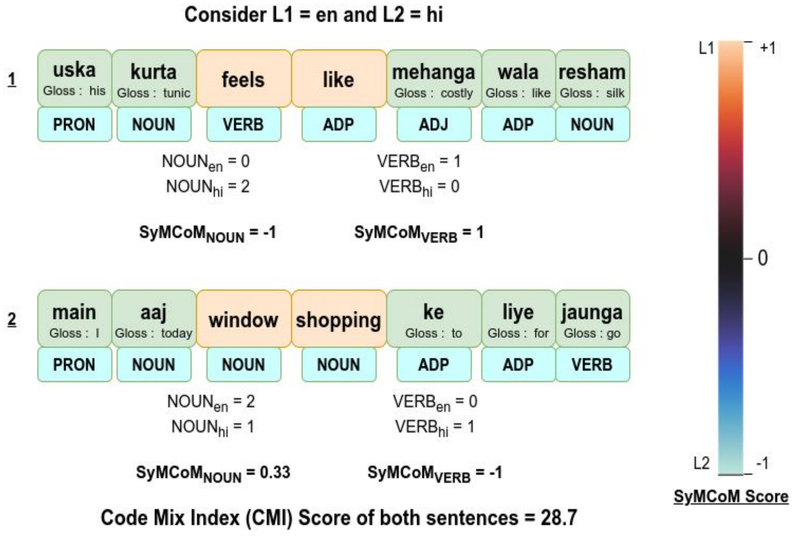
$$SyMCoM_{SU} = \frac{(Count_{SU_{L1}}) - (Count_{SU_{L2}})}{\sum_{i=1}^{2} Count_{SU_{Li}}} \quad (1)$$

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

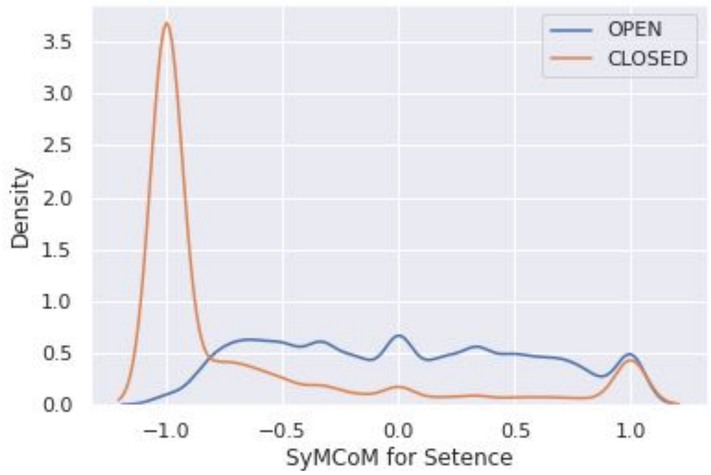
$$SyMCoM_{corpus} = \sum_{sent} \frac{SyMCoM_{sent}}{\#\ sentences\ in\ corpus} \quad (3)$$

- SyMCoM is a metric to measure syntactic variety in code-mixed text, at various levels - PoS tag, sentence, corpus.

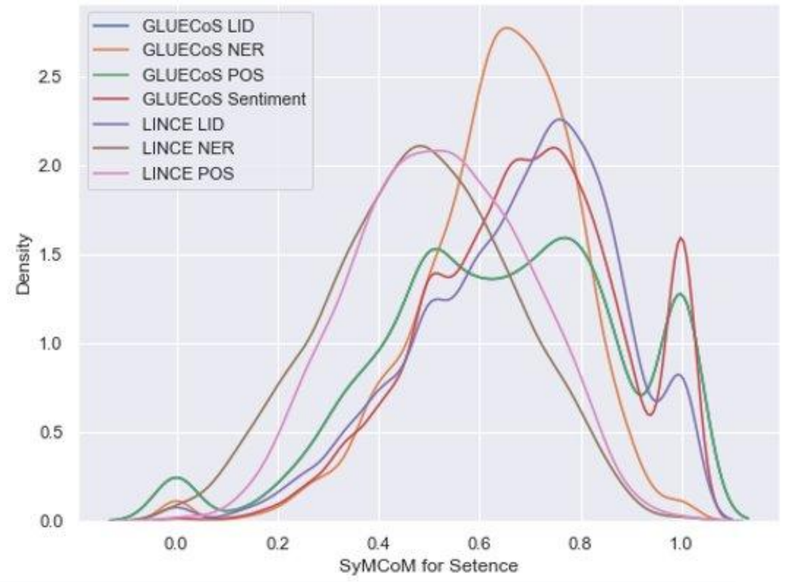
- Can distinguish between sentences where CMI would fail



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



ACL 2022, Dublin