Understanding Code Mixing: A Syntactic Perspective with SyMCoM

Himanshu Pal

October 2023

1 Introduction

Code mixing is a linguistic phenomenon of switching between languages in conversations and has been a subject of extensive study. However, understanding syntactic nuances in code-mixing and how they influence computational models and get influenced by them remains a relatively unexplored territory. This paper delves into proposing a novel metric system called SyMCoM, which stands for Syntactic Measure of Code Mixing. SyMCoM aims to quantify the syntactic variation in code-mixed text, particularly focusing on English-Hindi (en-hi) code-mixing.

2 Hypothesis and Experiment Setup

The paper hypothesized around the syntactic behavior of code-mixing, especially in terms of Parts-of-Speech (PoS) categories. The paper also found quantitative evidence that certain categories, such as CLOSED class categories and verbs, are less likely to be switched compared to others. **The only assumption that this paper makes is that SyMCoM can be computed for any corpora as long as there is a reasonably accurate POS tagger for the code-mixed language pair.**

2.1 Data Collection and Preprocessing

The study gathered datasets of English-Hindi code-mixed text, employed a pre-trained character-level BiLSTM Language ID tagger to determine the language for each word, and fine-tuned a state-of-the-art English-Hindi PoS tagger with an accuracy of 93.4% to accurately identify and label the Parts-of-Speech (PoS) in the corpus.

2.2 Calculating SyMCoM

A precise way to quantify variations in the syntax of code-mixing patterns should ideally encompass two crucial aspects: a) determining whether a Part-of-Speech

(PoS) tag or a syntactic category has been switched, and b) gauging the extent or contrast of this switch between Language 1 (L1) and Language 2 (L2) for that particular unit.

To encapsulate these attributes, it proposes a metric termed Syntactic Measure of Code Mixing (SyMCoMSU), formally defined by the equation:

$$SyMCoMSU = \frac{(CountSUL1) - (CountSUL2)}{\sum_{i=1}^{P} CountSULi} \quad (1)$$

The SyMCoM scores are bounded between -1 and 1, with the polarity indicating the language contributing more tokens. The SyMCoM scores for sentences are the weighted average of absolute SyMCoMSU scores for all SU, where the weights are the fraction of tokens in the sentence belonging to an SU. It is bounded between [0,1]. Values closer to zero indicate that L1 and L2 contribute nearly equally for most types of SUs in the sentence, whereas values close to 1 indicate that in the sentence, each SU is majorly contributed by a single language. Similarly, SyMCoM scores for sentences can be averaged over the corpus to capture the syntactic variation at a corpus level.

3 Results and Discussions

The result of the experiment unveils intriguing insights into code-mixing from a syntactic perspective. Analysis validates the fundamental tenets of the matrix language theory. highlighting that CLOSED class categories and finite verbs are less likely to undergo language switching in code-mixed text. This affirms the initial hypothesis and contributes to a deeper understanding of code-mixing patterns.

We observe that SyMCoM scores, derived from our novel metric, are significantly impacted by the accuracy of the PoS and LID taggers. Any errors in these taggers can introduce noise into the SyMCoM scores, emphasizing the importance of robust language identification and PoS tagging mechanisms.

4 Conclusion and Future Directions

In conclusion, the paper introduces SyMCoM, a syntactic measure tailored to analyze code-mixed corpora. It demonstrates its utility in understanding code-mixing, highlighting syntactic variations across different PoS categories. Findings in the paper shed light on the syntactic complexity of code-mixed text, emphasizing the roles of open and closed class categories. **The successful application of SyMCoM in English-Hindi code mixing prompts further research, including extending the metric to multilingual scenarios and deeper syntactic structures.**

Future work can explore extending SyMCoM to code-mixing involving more than two languages and incorporating deeper syntactic structures, such as nested phrases, to provide a comprehensive analysis of code-mixing complexity.