Tanish Lad (2018114005)

Jashn Arora (2018114006)

Automatic Generation of Code Mixed Sentences plus issues in Automatic Generation

Mentor Name: Pruthwik Mishra

Project Outline

Problem Statement

Given a Parallel Corpora (Sentence Aligned Corpora), the task is to Generate Synthetic Code Mixed Data.

Our Approach

After looking through Previous Works on this topic, we would be Combining Chunks of both the languages (Hindi and English) by using the outputs of Giza++, Stanford Full Parser (for English) and LTRC Chunker (for Hindi) and then matching head words found in the output of Giza++.

Deliverables till First Deadline

1. **Successfully Run Giza++ tool:** To find the word alignments between English and Hindi.
2. **Run Chunkers on Hindi and English Data**

Deliverables till Second Deadline

3. **Combining Chunks:** Matching the head words found in the output of the Giza++ software.
4. **Run Language Model (Optional):** To find out naturalness of code-mixed sentences.