Tanish Lad (2018114005)

Jashn Arora (2018114006)

Automatic Generation of Code Mixed Sentences plus issues in Automatic Generation

Mentor Name: Pruthwik Mishra

Interim Report

Work done upto 5/4/19:

1. **Ran Giza++ Tool:** We ran the Giza++ tool to find out the Word Alignments only to find out that the output will be of no use (As told by our Mentor) because of the problems in the Data Set.

2. **Filtered Data:** We filtered out the incomplete data in cases where No Hindi Translation was available corresponding to the English Data or No English Translation was available corresponding to the Hindi Data. We also removed some cases where the English Chunker was recognizing Mr. X Y as two different sentences (1: Mr. and 2. X Y) as told by our Mentor.

3. **Ran Chunkers on Hindi and English Data**: We successfully ran Chunkers on Hindi (LTRC Chunker) and English (Stanford

Dependency Parser) Data. Now we have the Chunked English and Hindi Data.

Work to be Done:

4. Firstly, we have to make the outputs of both the Chunkers in the same format.

5. **Run Giza++ tool Again:** To find out the Word Alignments of the Head Words and get a Translation Table (We changed the order of the steps so as to accommodate problems in English Chunker)

6. **Combining Chunks:** Matching the head words found in the output of the Giza++ Tool.