Automatic Generation of Code Mixed Sentences

Intermediary Project Report

Saransh Rajput-2018114016
Anshul Padhi-2018114013

Done so far :

1) From the given big dataset , we extracted usable components .

2) Align the Hindi and English corpora.

3)Run shallow parser on Hindi corpora .

4) Run Stanford full parser on English corpora to get

chunks in Hindi and English .

5) Run Giza++ from Hindi to English and English to Hindi.

Link to all our files
https://github.com/AnshulP10/CL-Project

Plan ahead :

1) On passing our corpora through the Stanford parser , the actual output was much more than expected output. This could be because some sentences are split by the parser, which leads to a greater number of sentences. This creates a huge issue as it disalignes our two corpora, thus would give the wrong result.
So we need to find the sentences which split when passing through the parser , fix or Delete these sentences.

2) Once the above problem is solved , remove unnecessary components from the parser outputs and then extract the necessary information(The pos tags and head of each chunk).

3) Shallow parser provides the head for Hindi chunks but the Stanford parser does not. So we need to formulate grammar rules in English to find the heads of the chunks.

4) Run Giza++ both ways again to get the probabilities . We can use the result to get more accurate results. If a word in Hindi maps to a given word in English when we run Giza++ from English to Hindi and the English word corresponds to another Hindi word when we run it from Hindi to English then we cannot use it as a correct replacement. But if two words(one in English and one in Hindi) map to each other when we run Giza++ both ways we can use them as a replacement for one another. Thus the quality of the generated text will be better.

5*) If time permits , we plan to come up with a language model which checks the authenticity of a code mixed sentence. Not every code mixed sentence seems natural. Like if said "Mai apple Khaa rha hu " it seems natural. But " Mai food Khaa Raha hu " does not seem natural. So we plan to come up with a language model which assigns a score to each sentence on the basis of its naturalness and only if the score is above a certain threshold, we call the sentence as a valid code mixed sentence.