Auto Generation of Code-Mixed Text

Project Outline

Anshul Padhi-2018114013
Saransh Rajput-2018114016

Project Description

Given two corpora, where the two are of different languages and one is the translation of the other, we have to generate code mixed sentences.

Project Outline

Steps

1.The given corpora may have characters which cannot be processed using the given tools. So we have to remove the sentences which have such characters and thus remove the corresponding sentences in the other language.
2.Align the sentences in the two corpora so that each sentence is mapped to its translation.
3.Align the words in the corpora for each sentence using Giza++.
4.Chunk the English corpora using the Stanford Full Parser.
5.Chunk the Hindi corpora using the Hindi chunker built by LTRC.
6.Extract the output in the desirable format so that we can further use it.
7. The output of the LTRC chunker gives the head, but the Stanford one does not, so we have to come up with rules to find the head of the chunk.
8.Now we map the heads of one corpus to the corresponding head of the other corpus and we replace the two only if word A is aligned to B and word B is

aligned to A. We can check this by running giza++ two ways.

Project Status

Till now we have :edited the corpora
aligned the corpora
chunked the corpora