CoMeT Adv NLP project

Interim submission

Anantha Lakhmi (2020101103)
Bhargavi Kurukunda (2020101077)
Nanditha Merugu (2020102061)

Preprocessing:

Here we passed an English file as an input to generate Hindi files for all test, train, and dev categories to use in mBARThien model.
We had used the below reference for the dataset:

Training:

> We basically split the words in both the english and hindi input files produced during the preprocessing stage.
> The output files depends on the MODEL used

1. If the MODEL used is mBARTen, then the output file consists only of english stripped data.
2. If the MODEL used is mBARThien, then the output file contains both hindi and english separated by a tokenizer(##).

> All the generated output files are stored in a newly created directory in the same folder under the name “data”.
> We had made sure to prune and fine tune the data
> Most of the words in the large vocabulary used by the original pre-training model are not actually used in the finetune process, so this part of redundant information can be removed. In the NMT model, the embedding matrix actually accounts for the largest proportion of parameters, so our cutting work mainly focuses on reducing the embedding matrix in the pre-trained model.