

Machine Translation

Lab 6

Master in Computational Linguistics – 2nd year
Academic year 2022-2023

Implement a phrase extractor for phrase-based machine translation.

The input files for your task are the 6 aligned files you already have from running the IBM algorithm in both directions on the three pairs of ro-en sentences, obtained either through GIZA++ or from your IBM implementation. Use those files to perform the following tasks:

1. Implement a phrase extractor able to extract all **consistent** phrases from your aligned files.
2. Perform symmetrization on each pair of sentences, to enhance the alignments. You will have lot of different situations which don't match the simple intersection/union rule, for which you will need to make explainable decisions.
3. Explain in a half page report your findings.