Between alignment and decoding, there exists the task of finding the phrase translations from the alignment to be used in the decoder. After all, the decoder needs a translation model in order to calculate translation probabilities.

You are given the first 10,000 lines of the Europarl parallel corpus to use as training, and your assignment is to extract the phrase pairs and their corresponding probabilities. You are also provided a French-English alignment for each sentence in the parallel corpus. Using this alignment, you will need to create a phrase translation table. Make sure to follow the format in the default implementation!

To evaluate the quality of these phrase extractions, we decode the testing data (1000 more lines of Europarl) using the extracted phrase table and an existing language model using the simple baseline decoder, and evaluate the resulting translation using BLEU. Thus, your task is to improve your phrase table in order to increase your BLEU score as much as possible.

The provided default currently creates phrase pairs by simply linking together words that are aligned together. There are several ways to improve upon this system. One way is to implement the phrase extraction algorithm described by the book. It focuses on finding “consistent” phrase pairs, and accounts for unaligned foreign words.

Of course, there are many other ways to improve the phrase pair extraction.

Here are some texts that further describe the problem:

1. http://acl.ldc.upenn.edu/N/N07/N07-2053.pdf
2. http://acl.ldc.upenn.edu/N/N03/N03-1017.pdf
3. http://mt-archive.info/MTS-2005-Vogel.pdf
4. http://maroo.cs.umass.edu/getpdf.php?id=187
5. 5.2.1 in textbook

To run the default, just simply run ./default and then run ./grade. Ensure that in your implementation, the translation model is written to data/tm and is consistent with the default implementation’s format.