Homework #3: CMPT-825

Anoop Sarkar – `anoop@cs.sfu.ca`

(1) The Canadian Hansards are the proceedings of the Canadian Parliament transcribed in English and French. The raw data has been cleaned up by Ulrich Germann (while at ISI) and provided for your use as a dataset of 1.3 million pairs of aligned French/English sentences.

The data is available at `/cs/natlang-a/data/isi-hansards/`. Please read the README file in that directory to appreciate the difficulty in creating the aligned sentence pairs from the raw text and exactly how it was done.

   a. Implement IBM Model 1 training on the Hansards (house and senate debates) training data.

   b. Implement the algorithm for extracting the best alignment given Model 1 parameters on the test data.

   c. How can we produce a English-French dictionary (mapping French words to English words) using Model 1 parameters?

   d. The Further Reading section on the course web page contains links to papers that extend Model 1 training in various ways. Try to implement the extensions, and try to come up with new simple extensions that improve the quality of alignments. Only use parameters $t(f_j \mid e_{a_j})$ and $\epsilon(m \mid l)$ to remain in the spirit of Model 1.