CSE573 - NLP Applications
Assignment 1: Word Alignment

Deadline
11:55pm, 5 February, 2020

Data:

The data for this assignment is available on the drive link here, which consists of sentence aligned English-Hindi, about 49k sentences. Use your IIIT email ID to access the data.

Task Description:

In this assignment, you are expected to implement the following word-alignment models:

1. IBM Model 1
2. The HMM model of Vogel et al. (1996)

The first probabilistic model to implement is IBM Model 1. You can find the details about the algorithm and more to implement this in the resources shared on moodle, and here.

The second is Hidden Markov Model. The details about this can be found in the paper - Stephan Vogel, Hermann Ney, Christoph Tillmann (1996), HMM-Based Word Alignment in Statistical Translation.

Here are some other references that might help:

- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer (1993) The Mathematics of Statistical Machine Translation
- Franz Josef Och, Hermann Ney (2003) A Systematic Comparison of Various Statistical Alignment Models
- Knight, Kevin, (1999) A Statistical MT Tutorial Workbook

Evaluation metrics should be Precision, Recall, and Alignment Error Rate (AER). You can find more information about these on NLTK's page here.

A 1-2 page write-up should be submitted along with the code, where the implementation choices have to be explained, comparison of the 2 models, and error analysis.

The results that are reported should be on the test data provided.

Language:

Python 3.x

Grading:

40 - IBM Model 1
50 - HMM
10 - Report
(+20) Bonus: You will be awarded bonus marks based on the level of innovation (use of extra features) that you bring to the experiment.

Note:

- You are expected to write your own code. Feel free to discuss the assignment with other students and collaborate on developing algorithms at a high level. Your report and the code that you submit must be yours. Plagiarism either from the internet or from other students will result in 0 on this assignment, and possibly more.
- Please make sure you start the work early, and submit it in time, keeping in mind the college festival and your quizzes for other courses that you might be taking.
- Use the moodle thread for any queries that you might have.