Design Document

Model1.py

This python file implements the IBM model1 and EM algorithm to translate the foreign to English sentences and print the alignments of the word according to the probabilities calculated by EM algorithm.

DATA STRUCTURES USED

Nested List- Used to store sentence of english and foreign language Pairwise.
LIst-Used to store english and foreign unique words.
Dictionary-It is used to store Probabilities of english word given foreign word.

INPUT GIVEN

Two files one containing list of english sentences and other containing foreign sentences.

OUTPUT GIVEN

It will print alignment of every french word corresponding to english word according to probabilities calculated by EM algorithm.
Time Taken -1.4 sec

Ibm.py

This python file implements python nltk library IBMModel1 and IBMModel2 implementations and compares the results with both models.

DATA STRUCTURES USED

Nested List- Used to store sentence of english and foreign language Pairwise.
LIst-Used to store english and foreign unique words.

INPUT GIVEN

Two files one containing list of english sentences and other containing foreign sentences.

OUTPUT GIVEN

It will print each pair of words along with their ibm model 1 and model 2 probabilities.

Time Taken -2.14 sec

phrase.py

This python file uses phrase-based translation model to translate french sentences into english sentences. It uses alignments obtained by our IBM Model 1 implementation as inputs. Phrase extraction algorithm extracts all consistent phrase pairs from a word-aligned sentence pair.

DATA STRUCTURES USED

Nested List- Used to store sentence of english and foreign language Pairwise.
LIst-Used to store english and foreign unique words.
Dictionary- It is used to store phrases obtained from phrased based translation module in nltk.

INPUT GIVEN

Two files one containing list of english sentences and other containing foreign sentences.

OUTPUT GIVEN

It will print phrase scores for each extracted phrase and rank them in order of descending probability.
Time Taken -1.86 sec