Assignment 3

Dibyendu Roy Chaudhuri

MT19034

Q1) HMM

The training data consists of around 15,711 POS-tagged sentences from the BERP corpus. The sentences are arranged as one word-tag pair per line with a blank line between sentences, words and tags are tab-separated. Contractions are split out into separate tokens with separate tags. Around 36 tags are present.

Assumption:

For Out of vocab words, I have used add-1 smoothing in instead of normal probability function in viterbi algorithm.

Algorithm working fine up to a satisfactory mark.

Q2) MEMM

There are a total of around 9751 unique words and

52 unique tags in the corpus. Most common tag present in the corporse is 'NN'.

Each line constitutes the word, POS-tag and Noun group tag. The noun

group tags are among: B-NP, I-NP and O

- B-NP – indicates a token begins a Noun Group
- I-NP – indicates a token is inside a Noun Group
- O – indicates a token is outside of a Noun Group

Accuracy: 80.60 %

Assumption:

For Out of vocab words, I have used add-1 smoothing in instead of normal probability function in viterbi algorithm.

Algorithm working fine up to a satisfactory mark.