Report for Assignment 3

Yisha Wu

December 18, 2013

I have designed a IOB tagger using Viterbi algorithm. The training set is provided in the course website with more than 400000 lines or 13000 sentences of IOB-tagged sentence corpus.

I assume all the words that only appeared one time in the training set are treated as 'UNKOWN' and all the words in the test set that never have been seen in the training set are also treated as 'UNKOWN'. And I also have done the smoothing for the tag transition probability.

For the 'UNKOWN' words, I declared some shape features. If the length is more than 8, the word is 'UNKOWN_long'; if the word string has numbers, the word is 'UNKOWN_num'; if the word string has both upper case and lower case, it is 'UNKOWN_mix' and finally if the word string is full of upper case, it is 'UNKOWN_upper'. And all the other 'UNKOWN' words are seen as 'UNKOWN_others'.

Since the tags in this assignment are only 'I', 'O' and 'B', I didn't count the number of tags. Instead I just count the number of different words in each tag and compute the word probability of each tag.

The inputs of my Viterbi algorithm are: the tuple of all words, the tuple of all tags, the start probability dictionary about $P(t_i|start\_tag)$, the probability dictionary about $P(t_i|t_{i-1})$ and the probability dictionary about $P(w_i|t_i)$.

The output of the my Viterbi algorithm is the most likely POS tag sequence of the given sentence. The output file format of my system is the same as the format of the training set file.

I use 90% of the training set as the training data and the remaining 10% as the test data to get the accuracy of my Viterbi algorithm. And the F-measure is about 60%, which is pretty good. So I believe that my system works well.