Jack Cardwell

Professor Bailey-Kellogg

Computer Science 10: Problem Set 5

May, 20, 2015

Discussion of the Effectiveness of HMM and Viterbi

A highway department source said there also is a plan there to issue some $3 million to $4 million worth of rural road authority bonds for rural road construction work.

Brown tag: DET N N N VD EX ADV V DET N ADV TO V DET N NUM P N NUM N P ADJ N N N P ADJ N N N .

Viterbi Tag: DET N N N VD ADV ADV V DET N ADV TO V DET N NUM P N NUM N P ADJ N N N P ADJ N N N .

Attorneys for the mayor said that an amicable property settlement has been agreed on

Brown tag: N P DET N VD CNJ DET ADJ N N V V VN ADV .

Viterbi tag: N P DET N VD CNJ DET ADJ N N V V VN ADV .

In the places where the Viterbi algorithm went wrong, words must have been tagged as more than one parts of speech, where the tag that the Viterbi algorithm placed must have occurred more than the correct one in the training files. Although more training data would’ve increased the chances of the algorithm getting the correct tag, there is always a chance that a word can be tagged incorrectly if it can be used as more than one part of speech. Overall, this algorithm did a fairly good job determining the parts of speech. Specifically it was 86% correct for the simple test cases and 96.51% correct for the larger, brown data. Evidently, training with a larger data set greatly improves the efficiency of the program. The smaller the unseen word probability, the greater number of mistakes that the algorithm made. For example, when set to -10, the function got 3989 words incorrect, versus the 1271 incorrect with an unseen probability of -100. By making the number more negative, it decreases the chances that a word, unseen to a specific state can be selected as the emission from that state. The unseen word probability most definitely plays an important role in the method because the concept of Viterbi is applying a training dataset of tags to words that are never seen. If a word is unseen, the transition component and the previous component will have greater importance because they are the only two things that the dataset provides, when the word is not recognized.