Ben Williams

Alex Craig

PS-5 Writeup

Here is one example of where our bigram model worked perfectly:

*meanwhile , it was learned the state highway department is very near being ready to issue the first $30 million worth of highway reconstruction bonds .*

This was one of the first sentences in Brown Test Sentences (#2).

However, here is one where our model missed a term:

*sheets added that he would ask for exclusive use of voting machines in the state by january , 1964 .*

The model misclassified “sheets” as a noun, and not a proper noun (brown test sentences #16).

This is understandable. “sheets” is much more likely to be used as a noun than a proper noun in most texts, and sentences also often start with nouns. Therefore, the model really had no reason to think that “sheets” would be a proper noun over a noun.

The biggest factor in believing that our model was correct is that it had a high success rate over the Brown Test Sentences (94.69%). If something were fundamentally wrong, it would have shown itself over this large scale test. Early on, using the given simple train/test sentences helped us know we were going in the right direction. Also, this percentage value was not very dependent on our non-observation penalty, meaning that the bulk of the correctness came from the POS->POS probabilities.

Extra Credit

For extra credit, we implemented a Trigram model alongside our Bigram model. You can choose which one you want to use in the scanner/input. We did this by also keeping track of the POS two words ago during training, and then calculating the probability of the newest POS given the previous two POS’s. Then, in the viterbi, we could start taking into account these probabilities when we were at the 2nd word of a sentence. We had to add a check to make sure that the HashMap used was not null, as periods often had no HashMap given that they are always at the end of the sentence. We found that this barely increased our accuracy (only 0.02 %), which we found surprising given that you would think more information leads to more accuracy. Perhaps a further step would be to add another layer of word observations two ahead, which could be implemented very similarly to the normal word observations, but given that the word penalty was generally very negligible anyway, this may not make a difference.