Josh Weeks
reportpart2.pdf

I found it interesting that some of the lower frequencies of sequences contained some information that I would not call a verb phrase such as 'NN CC NN' that only has a frequency count of 1. Noun, coordinating conjunction followed by another noun should not be, in my understanding, a verb phrase. I also found that most verb phrases start with some for of a verb but not always. Modals seem to be quite common in their occurrence before a verb. This was the third highest frequency tag sequence that was found in my analysis. In the top 30 there seems to be a lot of adverbs occurring also. VBD is the highest frequency sequence by itself at a count of 5280. This is where I made progress in making a better chunker.

My analysis of the tag sequences showed me that the most common verb phrase in the training data is VBD, which is a verb in the past tense. I used this to develop a function to include in my feature extractor. My feature extractor included lookahead features, paired features, and complex contextual features. The complex contextual features come from my function tags_since_vbd. This feature creates a string describing the set of all parts of speech that have been encountered since the verb in past tense. This had a greater ChunkParser score than tags_since_vbz, which is the same idea as tags_since_vbd but with VBZ (a 3rd person singular present verb). tags_since_startwithvb was an attempt to make a better score than the tags since VBD but I was not successful. I thought that if it extracted sentences by using pos tags that begin with VB, it would improve since I am looking at verb phrases. I tried to combine the three function and got a much lower score. So my next thought was to just add more complex contextual features to my feature extractor for each function created and this also gave me lower scores.

My conclusion from this is that the chunker was better when the feature set was based off of the tags that occur since a VBD. This made the classifier perform better than without any complex contextual feature in all scoring except IOB accuracy. Precison,recall, and F-measure increased by 0.1%. This is a good increase since we are dealing with a large set of data.

NOTE: In my ConsecutiveVPChunkTagger, I am using the Naive Bayes classifier to produce faster results.