Ryan Richardson

NLP

HW3

1. There are 2182 total usages of the word to, with most of them being categorised as ‘TO’. There is one instance of the word ‘TO’ where it is marked as a proper noun. In that instance, it was a title of a response to the defense counsel from the prosecution.
2. The most common incorrect tag is punctuation, where the % symbol is marked as a noun 647 times in the training set and 74 times in the test set. I only noticed a few other incorrect NN tags unless the tags should have been considered proper nouns, in which case ACCOUNT and Pressure were used a few times.
3. Based on the Penn-Treebank tagger, we can add a few basic rules and lists that will be used for tagging, such as further breakdowns for types of adjectives and adverbs. We can consider changing the order of our patterns, as the first matched pattern will be the tag. As a result, we should ensure that the ‘special’ rules for Adjectives and Adverbs follow their respective base rules for an adjective or adverb.   
     
   I considered changing the default tag to ‘None’, as tagging the remaining untagged words as Nouns may decrease accuracy. however, if we have our other rules properly, the only untagged words should be nouns or proper nouns.  
     
   Overall, I added a number of rules, especially those involving word lists, Modal Verbs, Proper Nouns, Personal and Possessive Personal nouns, Comparator and Superlative Adjectives and Adverbs, Determiners, Interjections, and Coordinating Conjunctions.  
     
   Even with these updates, the tagger was still only 10% accurate, compared to the unigram tagger’s 45% and baseline taggers of 58%. though this may be partially down to using different tags for the parts of speech based on the Penn-Treebank tags instead of the default NLTK tags.
4. By default, our Brill tagger has an accuracy of 73.3%, and the baseline tagger has an accuracy of 72.9%.
   1. We can increase the baseline tagger by combining multiple taggers on top of each other. At first, I simply tried using Unigram, Bigram, and Trigram taggers on top of each other and got a small accuracy increase to 73.1%, consistent but hardly significant. This reinforced that we don’t continually increase accuracy by adding more context.   
        
      However, by chaining two affix taggers, we saw our accuracy increase to 80.31% when looking at the prefixes of words. This makes sense as it means that we are looking at the core word to generate a tag that can make it more accurate. Think of plural words, adjectives, and adverbs that can be impacted based on their suffixes, so by choosing to put more weight on the prefix of a word, we can ensure proper tagging.   
        
      This also increases the accuracy of our Brill tagger to 80.38% accuracy. We did not add a Bigram tagger to our combined tagger as the context from the Trigram tagger was enough, and the context gained from the Bigram tagger was reducing our accuracy by as much as 5%, so it was not used.
   2. I changed the Brill tagger logic to look at an expanding range around a target word and tag. While we have previously found that adding limited contextual clues did not help with our accuracy scores, the hope is that the brill tagger’s method of building its tagging logic from the provided templates would be sufficient in building additional context without compromising the base tags of the corpus. Unfortunately, despite trying a wide range of different templates and methods, I was unable to get the accuracy to improve by more than ~0.01%, which is still at least partially significant on a large corpus but not as significant as I would hope.
   3. Rule 352 changed the word ‘attorney’ from an adjective to a noun, using the template Word[-1]
5. Using the Brill Tagger on the treebank corpora we get an accuracy of 69.38% indicating that the tagger may have been overfit to the brown corpora and would need some additional training and tweaking in order to be better suited as a general tagger.