6.863J/9.611J: Natural Language Processing Prof. Robert C. Berwick

Assignment 4: HMM named-entity tagging

My posse consists of: Sam Ainsley *Salman Ahmad (saahmad@mit.edu)*

1. **Problem 1**

   The results from the unigram are:

```
Found 14043 NEs. Expected 5931 NEs; Correct: 3117.

 precision  recall  F1-Score
Total:  0.221961 0.525544 0.312106
PER:  0.435451 0.231230 0.302061
ORG:  0.475936 0.399103 0.434146
LOC:  0.147750 0.870229 0.252612
MISC:  0.491689 0.610206 0.544574
```

2. **Problem 2**

   The results from the bigram are:

```
Found 4472 NEs. Expected 5931 NEs; Correct: 3128.

 precision  recall  F1-Score
Total:  0.699463 0.527398 0.601365
PER:  0.617253 0.400979 0.486148
ORG:  0.531476 0.384903 0.446467
LOC:  0.841415 0.700109 0.764286
MISC:  0.756066 0.642780 0.694836
```

   The results seem to be an improvement over the unigram model.

3. **Problem 3**

   The results from the trigram are:

```
Found 3926 NEs. Expected 5931 NEs; Correct: 3270.

 precision  recall  F1-Score
Total:  0.832909 0.551340 0.663488
PER:  0.861290 0.435800 0.578757
ORG:  0.712644 0.417040 0.526167
LOC:  0.860634 0.710469 0.778375
MISC:  0.869814 0.660152 0.750617
```

   As can be seen, the results improved from both the unigram and bigram models.

   The one thing that stood out in development is that you needed to rely on smoothing to get decent results. Computing the trigram probabilities will often result in a bigram in the denominator that has never been seen before. For this problem, I just implemented Laplace smoothing to my trigrams and that worked well enough.

4. **Problem 4**

   The results for my final tagger are:

   `TODO`

   The first thing that I explored was increase the frequency counts.

   I created a class for numbers as those should almost always be "O". They were added to a class called `_NUMBER_`

   I created a class for punctuation. They were added to a class called `_PUNCTUATION_`

   I created a class for capital abbreviations (like "M.") as these are almost always I-PER. They were added to a class called `_ABBREVIATION_`

   The rest fell through and were added to `_RARE_`

   All caps

   Capital, the rest lower case

   Using word net?