burkhardt-amy-assgn4-report

November 30, 2017

1 A Hidding Markov Model Approach for the Named Entity Recognition task of identifying genes in biomedical journal article abstracts.

1.1 Overview

My approach to this assignment, which was to "implement a learning-based appraoch to named entity recogntion (NER)" is almost identicial to the HMM sequencing labeling task (using a Viterbi decoder) of the second homework assignment -- but instead of hidden states of part-of-speech tags, in this example, the hidden states are IOB tags.

1.1.1 IOB Tagging

In IOB tagging for this work, the begining of a gene is labeled with a (B), all subsequent tokens of the gene are labeled with an (I) indicating that they are inside the entity type, and an (O) is assigned to all the tokens outside of the gene entity.

1.2 Intermediate Results

Although this present implementation handles the decisions around the smoothing and treatment of unknown words in an identical manner as the first HMM assignment, a handful of different parameters were tested, and evaluated based on the intermediate validation sample (DevTest), which was a random 20% of sentences within the gene-trainF17.txt file (the other 80% of the data was used to train the model).

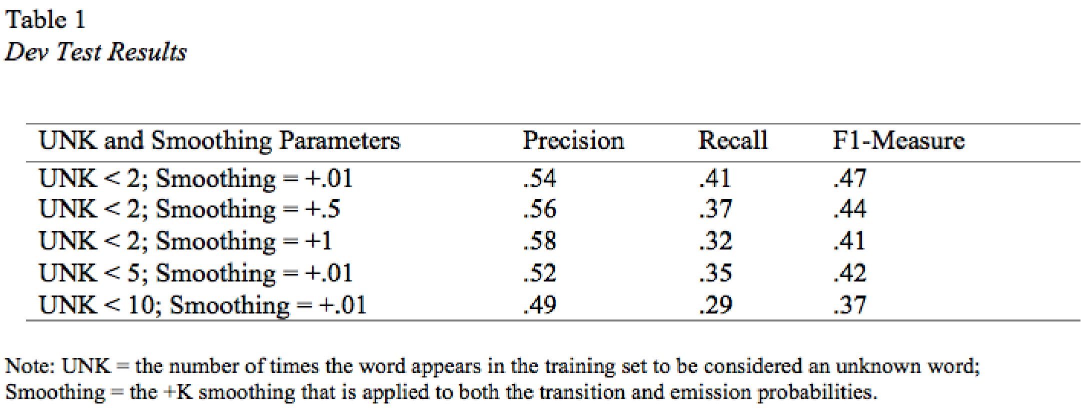

Table 1
Dev Test Results

UNK and Smoothing Parameters	Precision	Recall	F1-Measure
UNK < 2; Smoothing = +.01	.54	.41	.47
UNK < 2; Smoothing = +.5	.56	.37	.44
UNK < 2; Smoothing = +1	.58	.32	.41
UNK < 5; Smoothing = +.01	.52	.35	.42
UNK < 10; Smoothing = +.01	.49	.29	.37

Note: UNK = the number of times the word appears in the training set to be considered an unknown word; Smoothing = the +K smoothing that is applied to both the transition and emission probabilities.

For this assignment, we are most interested in the F1-Measure, which is the harmonic mean (Precision*Recall)/(Precision + Recall). As a reminder, both Precision and Recall have the number of true positives as the numerator and denominator. The difference is that the denominator for Precision is the sum of TP + FP; the denominator Recall is the sum of TP + FN.

In this excersise, a true positive is considered a correctly identified gene.

1.3 Possible Next Steps

This named entity recognition task used the sequence classifier of a Hidden Markov Model. However, a Maximum Entropy Markov Model (MEMM) might serve to improve the Precision and Recall. Using an MEMM would allow us to have both the IOB tagging, as well as other features that could be useful in determining whether or not a word or sequence of words is a gene. Some other features that I could use could be: if the token is a non-alphanumeric characeter (e.g., "-"), if the token contains both characters and numbers, if the token is upper-case, or if the token doesn't contain any vowels (e.g., Sp1, mRNA).