Final Project Pitch

Bill Xia

CS 136: Statistical Pattern Recognition, Tufts University

1 Goal

The goal of my project is to build a probabilistic model for identifying non-consumer biomedical terms (which I'll call expert terms going forward) in biomedical paper abstracts.

2 Data

My data was derived from the PLABA dataset, a collection of biomedical paper abstracts. Last summer, I annotated the data by constructing Beginning-Inside-Outside label sequences for each sentence. That is, each word in each sentence was labeled either B (beginning of an expert term), I (inside an expert term), or O (outside an expert term). The following is an example of such a labeling:

```
Quinine sulfate is an antimalarial drug.
B       I       O  O  O            O
```

Put together, the data I'll be training my models on is made up of sentences (strings) paired with lists of labels, where each label corresponds to a single word in a sentence.

3 Baseline

For my baseline, I plan on making a Markov assumption and predicting each word-label pair according to the pair immediately preceding it. So, if we denote *w* as a word and *l* as a label, we can write out the formula for my baseline:

$$\begin{aligned} p(w_t, l_t \mid w_{t-1}, l_{t-1}) &= p(w_t \mid w_{t-1}, l_{t-1}) * p(l_t \mid w_t, w_{t-1}, l_{t-1}) \\ &= p(w_t \mid w_{t-1}, l_{t-1}) * p(l_t \mid w_t, l_{t-1}) \end{aligned}$$

In the second line, I remove w_{t-1} to reduce the complexity of the right hand side. I will define parameters π and η to represent w_{t-1}, l_{t-1} and w_t, l_{t-1} respectively. They will be used in Categorical pdfs to model the probability of word label pairs in my baseline.

4 Upgrade

For my upgrade, I plan on implementing a Hidden Markov Model. Using a hidden variable (z) with length equal to the word and label sequences, I will be able to predict w_t and l_t using the full context of the sentence without my computational costs skyrocketing. The math will look similar to the baseline, but instead of looking only one step back, we'll be basing each word-label pair on the entire rest of the sentence via the hidden variable.