Aditya Agashe (ava9), Rohit Biswas (rb625), John Hughes (jlh457)
Kaggle Team Name: Hairy Macho Men (HMM)

Project 2: Part 1 Proposal

Baseline

First, we assembled a lexicon of uncertain tokens. By going through each token in the training files, we identified the ***n*** most common tokens marked in an uncertain phrase. We used this lexicon to identify phrases and sentences. To find the ranges and sentences in our test data, we first tagged all tokens that were in our lexicon and in our test data. To identify phrases, we emitted ranges of uncertain tokens in our test data into a csv. To identify uncertain sentences, we determined the density of uncertain tokens in each sentence. If the the density (number of uncertain tokens divided by number of total tokens) was higher than a threshold ***p***. Those sentence indices were then outputted into the csv. The parameters ***n*** and ***p*** were tweaked to generate higher scores in the kaggle baseline submissions.

Proposal

For our model, we plan to use an existing toolkit to implement CRF sequence tagging. We chose to implement this model because it allows us to define our own functions and this flexibility will allow us to fine-tune the model so it has the highest predictive power. We did not use HMM because it is very reliant on the entire sequence rather than the neighboring tokens, while we hypothesize that uncertainty phrase identification is more reliant on small phrases. We have come up with a couple different feature functions which we hypothesize should have predictive power at identifying uncertain phrases. They include (and may be expanded on):

- A phrase/sentence is automatically marked as uncertain if it contains a word in a manually defined dictionary (likely small size) that are likely to be uncertain
- Mark a phrase/sentence as uncertain if it contains more than ***x*** consecutive **tokens** that have non-zero uncertainty probability
- Mark a phrase/sentence as uncertain if it contains more than ***x*** consecutive **tags** that have non-zero uncertainty probability
- Mark a sentence as uncertain if it has a uncertainty density (as described above) higher than a certain threshold (could be determined by a validation set)

We will run experiments to try to determine which combination of these feature functions along with the appropriate parameters is most effective.

Extension

For the main part of our project we will be testing the effectiveness of different combinations of the feature functions. Since we will also be using a token sequencing library, we plan to test the effectiveness of different token sequencing methods.