Proposal of Project 3: Sequence Tagging

Shibo Zang(sz428) Hongfei Li(hl963)

We decided to implement a Hidden Markov Model for the task of this project -- Named Entity Recognition, and use Viterbi algorithm as the decoding algorithm.

There are three key points of our algorithm: feature selection, interpolation, and low-frequence word smoothing.

Since HMM is based on Bayes' theoreom, it is very important to choose appropriate features for the model. There are several common feature categories that we thought may be useful, like shape, character affixes, predictive tokens, lexical items, or syntactic chunk labels. For example, there are several types of shape features, like if the word is capitalized, if only the first letter is capitalized, if the word is mixed with lower-case and upper-case letters, if the word ends in digit, or if the word contains hyphen. For character affixes, we may find out all the affixes that is labeled as named entities and using them as one of the features.

For Hidden Markov Model, there are several options like unigram HMM, bigram HMM, trigram HMM, etc. Since we don't know the performance of these models at present, we could implement interpolation method while calculating the transmission probabilities of Hidden Morkov Model. For example, while calculating $P(\text{I-ORG} \mid \text{O, B-ORG})$, we could use interpolation method like $P(\text{I-ORG} \mid \text{O, B-ORG}) = \lambda_1 * \frac{Count(\text{O, B-ORG, I-ORG})}{Count(\text{O, B-ORG})} + \lambda_2 * \frac{Count(\text{B-ORG, I-ORG})}{Count(\text{B-ORG})}$ $+\lambda_3 * \frac{Count(\text{I-ORG})}{Count(\text{all tokens})}$, where $\lambda_1 + \lambda_2 + \lambda_3 = 1$. With interpolation method, we could conduct experiment on choosing the proper lambda values.

Besides, we need to implement low-frequency word smoothing to avoid zero probabilities. We would split vocabularies into two sets: frequency words, which occur more or equal than five times in the training corpus, and low frequency words, which occur less than five times.

And we plan to implement MEMM in addition to HMM and conduct experiments on different smoothing method.