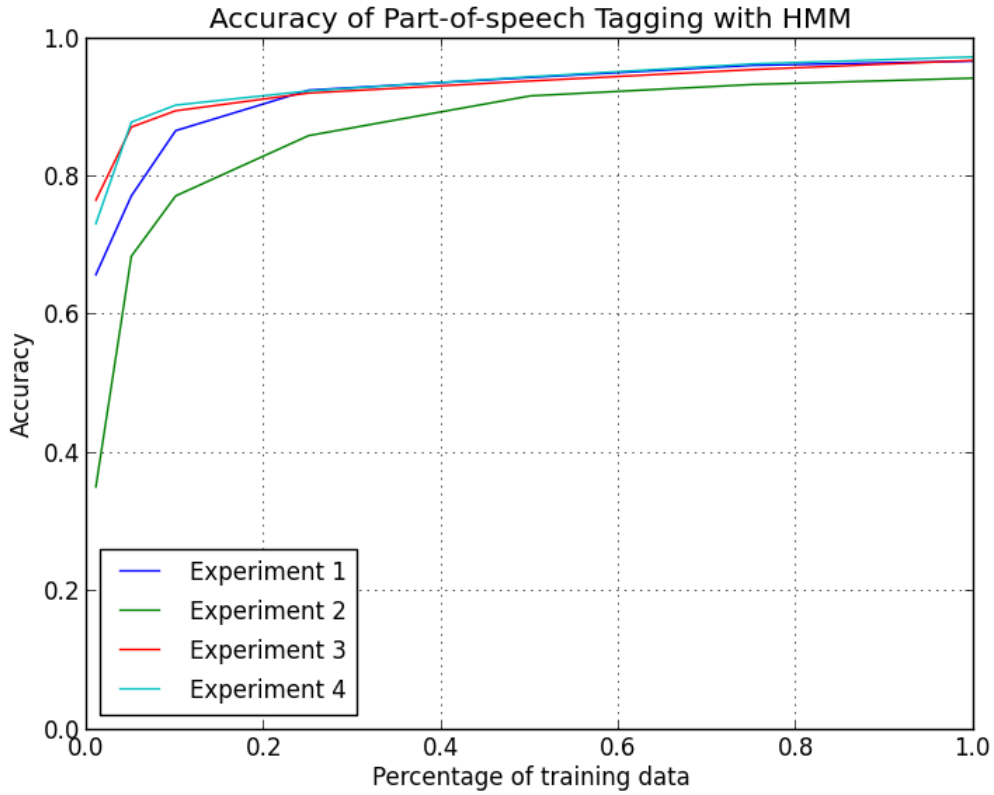
COMP 182 HW6 Problem 4

Yanjun Chen

Apr 21th 2017



From the accuracy curve in the figure, we could firstly notice that using smoothing method in our transitional probability produces better results than not using smoothing method. Also, after using smoothing, both digram and trigram models seem to differ little from each other. This may be caused by the other two terms that smoothing takes account into when calculating the transitional probability. What smoothing does is to "smooth" the strict causal relationship of Markov Chain by adding terms that reduce coincidences or extreme cases in the data. For example, we add some bigram factors into our trigram model to reduce the second-order causality, and also consider the frequency that certain tags appear in our data to reduce the influence of extreme cases. Doing smoothing can make up the biases of our data, so no wonder it will give out better accuracy on our test.

Secondly, comparing bigram model with trigram model, we could see that bigram model works better than trigram model. It is usually presumed that second-order Markov chain is better than first-order. However, our experiments give out the opposite result, especially for the case of trigram without smoothing, whose accuracy is much lower than the other three. This may be caused by the nature of English language: part of speech in grammar seem to have more correlation with the prior term rather than the prior two terms. Our language is mostly consisted of two-word phases, like nouns usually go after verbs and verbs go after personal pronoun and so on. On the contrary, considering a three-word

phrases in our computation seems to deviate the causal relationship in our language, so we will instead get a lower accuracy when using trigram model. Also, a trigram relationship is somehow too strict, so extreme cases will more likely appear and deviate our results. The reason that trigram with smoothing works well exactly illustrates this logic. With smoothing, the relationship is smoothed and will thus render better accuracy.

Thirdly, if we focus on the same experiment, we will easily find out that the relationship between accuracy and percentage of data being used is kind of dramatic at first and gentle when the percentage goes up. This makes sense because when we use little of the training data, we will have a lot of unseen words that have no tags. As a result, we will not likely to find out which part of speech is the words. As the percentage of training data increases to some extent, we get most of the words we want so the accuracy curve states to go up gently. This is because we have already figured out the basic emission probability for the majority of the words, so keep adding data will only refine our emission probability but it won't have critical influence on our emission matrix.