CSC 248/448 Assignment #2
Due: Oct 2, 2003
CSC448 Students: Read Chapters 1, 2 and 6 of Manning and Schutze

Part I: Written Questions – Please hand in typed answers

1. Find at least one technical error in the lecture notes.

2. Consider a language generated by a sequence of random variables with 2000 different word types. We have a corpus of 1 million and one word tokens (1 million bigrams) and are interested in defining a good bigram probability model.

a) How many possible types are there in the bigram model of this language?

b) What is the maximum number of different bigram types that can occur in the corpus. What does this say about the lower bound on the number of bigrams that have not been seen.

c) Assuming there are actually 300,000 distinct bigrams in the corpus, what is the probability estimate for the next observed element to be a bigram we have not seen before, according to the MLE, Laplace and ELE estimators?

d) Again assuming the corpus contained 300,000 distinct bigrams. We empirically test a million new words of data and find that unseen bigrams occur 5% of the time. What value of λ in a Lidstone's estimatation would best match the data?

3. Prove that the probability function that maximizes the likelihood of a corpus is the same as the probability function that minimizes the log probability ("cross entropy") of the corpus.

4. Consider the following corpus from the ABC language
A B C B A B B C A A B C B C A B C
We want to build a good bigram model of this language using the linear interpolation technique

$$P_{LI}(x \mid y) = \lambda P_{MLR}(x \mid y) + (1 - \lambda) P_{MLE}(x)$$

We assume we have enough data that we can use the MLE technique for the individual distributions.

a. Determine the MLE bigram and unigram probability models.
b. Say we have a development corpus as follows: A C B C A C B C. What is the likelihood of this corpus for the following values of λ: 1, .9 and .5. Which appears to be the best estimate?
c. Find a better value of λ than the ones provided, or if you can't, given some evidence why this seems to be the case.
d. (448 students) We want to find a good (ideally optimal) value for λ. Is there an analytical solution? If so, derive it. Or if you choose to estimate it with an algorithm, sketch the algorithm.

Part II. Programming Assignment

We want to build a program that can identify the language from just a few words. We assume we do not have large vocabulary lists for each language and rather must perform our task by looking at letter sequence models computed from relatively small corpora for each language. You will be provided with five small corpora for five different languages.
a) Write a Perl program that can computes unigram, bigram and trigram letter probabilities from the training corpus using the Lidstone estimator, where the parameter lambda is an argument to your function. Note that if you set lambda=0, then your program computes the Maximum Likelihood Estimate (MLE). Design an experiment using the first corpus only to identify as effective value of lambda that maximizes the likelihood of new data for each of the n-gram models. Make sure you documentation carefully describes the design of your experimental method, the results you obtain, and the conclusions you draw.
b) Implement another estimation technique that was described in the notes that you think may produce better results that the add-lambda estimators. Describe your reasons for choosing this technique in your documentation. Perform some experiments that compares the performance of your new estimator with your best add-lambda model from part (a).
c) Using whatever models and estimation techniques you want, build five language models, one for each language, and then build a program that takes a sample of text and returns the most likely language it is written in. This program will be tested on a series of different test cases after you submit it. There will be evaluations in several categories: single words, word pairs and short sentences. Make sure you follow the instructions below so that your program can be run and compared with the others.

Submissions guidelines
Your program from part (c) will be tested semi-automatically. To make it easier, please name it identifyLanguage.pl. It should take a single file name with the text sample as an argument, and print the answer on the standard output. The answer must be one word: English, French, German, Portuguese, or Italian. A sample program invocation would be

```
linux:> identifyLanguage.pl english_test_file
English
```

Your program may print debugging and other information before the language identification, but the result must be the last word printed on the output, so that it can be easily extracted automatically. The code must assume that the file with the language model is in the current directory and contain no absolute paths.

Your submission should include all the program files in perl, your language models, a README file with the list of all submitted files and instructions how to run your programs, and the documentation file in PS or PDF format describing the program design, the techniques you used, and your conclusions. Please do not re-submit the corpora with your assignment.

Use the turn-in script to submit the assignment. On the CS graduate network, run

/u/myros/commands/TURN_IN .

On CS undergraduate network
/u/cs248/bin/TURN_IN .

As with previous assignment, make sure that your Perl code is well commented and easy to read and understand.