Mehryar Mohri
Speech Recognition
Courant Institute of Mathematical Sciences
Homework assignment 3
November 15, 2012
Due: November 30, 2012

A. n-gram models

For most of the questions of this homework, it is recommended that you use the GRM library and FSM or OpenFst libraries. In fact, try as much as possible to use the utilitites of these libraries to answer the questions. However, you need to justify your responses and not just mention the library utilities used.

1. Download the following training corpus S and test corpus $\hat{S}$:
   `http://www.cs.nyu.edu/~mohri/asr12/train.txt`
   `http://www.cs.nyu.edu/~mohri/asr12/test.txt`.
2. Extract the vocabulary Σ_1 of S and define a start and end symbol.
3. Create the following language models:
   - bigram back-off model;
   - trigram back-off model;

   Report for each of the weighted automata obtained
   - the number of states;
   - the number of transitions;
   - the number of ϵ-transitions;
   - the number of n-grams found ($n = 2$ for bigram models, $n = 3$ for trigram models).
4. Randomly generate 100 sequences from the first model and compare the likelihood given by the two models to the sample formed by these sentences.
5. Compute the perplexity of these models using the test corpus.
6. Shrink both of these models with the option $-s4$. What are the perplexity estimates for these models.

7. Create a simple class-based model where each class is either reduced to a single word, or a single bigram. To do so, use the average mutual information of $\Pr[w_1 w_2]$ and $\Pr[w_1]\Pr[w_2]$ to determine classes. Thus, a bigram $w_1 w_2$ forms a class of his own when

$$I(w_1, w_2) = \log \frac{\Pr[w_1\, w_2]}{\Pr[w_1]\Pr[w_2]} \tag{1}$$

is positive and relatively large.

   (a) List the ten bigrams $w_1 w_2$ in the training set with the largest $I(w_1, w_2)$.

   (b) Define the 2,000 bigrams with the largest $I(w_1, w_2)$ as bigram classes and create a class-based bigram back-off model (describe how the class-based model is defined using transducers and mention which FSM or OpenFst library commands were used).

   (c) Compute the perplexity of your class-based model and compare it with the previous perplexity measures obtained.

Maxent models

1. Download the Maxent software tools from:

   `http://goo.gl/E4EyV`

2. Use the software to train a Maxent model with n-gram features on the same training sample as in Problem A, for $n = 2$ and $n = 3$.

3. Compute the perplexity of each of the models obtained on the test sample (same as in Problem A). Compare the values obtained with those obtained for n-gram models in Problem A.