Mehryar Mohri
Speech Recognition
Courant Institute of Mathematical Sciences
Homework assignment 2
Due: November 21, 2007

A. N-gram Language Models

For most of the questions of this section, it is recommended that you use the GRM library. 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/asr08/train.txt`
   `http://www.cs.nyu.edu/~mohri/asr08/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;
   - bigram interpolated model;
   - trigram interpolated model.

   Indicate 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 10 sequences from the first, and 10 from the second model (you can use `fsmrandgen`). Which seem closer to English sentences?

5. Assume that the n-gram language model you derived is represented as a weighted automaton A over the log semiring. Describe an algorithm to compute efficiently the perplexity of A on $\hat{S}$ by constructing an automaton B representing the sentences of $\hat{S}$. What would be the sequence of fsm library or Openfst utilities used.

6. Estimate the perplexity of the first and fourth models on the training corpus S and on $\hat{S}$.

7. Shrink both of these models with the option $-s4$. What are the perplexity estimates for these models.

8. To come up with a model with lower perplexity than both p and q, Prof. Plexy suggests using a convex combination, $\alpha p + (1 - \alpha)q$, for some $\alpha \in [0, 1]$. Show that the perplexity of Prof. Plexy's solution is upper bounded by $\alpha\pi(p) + (1 - \alpha)\pi(q)$, where $\pi(r)$ denotes the perplexity of model r.