Mehryar Mohri
Speech Recognition
Courant Institute of Mathematical Sciences
Project – December 02, 2012
Project report due: December 14, 2012
Project presentation: December 17, 2012

The objective of this project is to learn a weighted transducer for spelling correction.

Description

The project consists of the following general steps.

1. download the following training and test sets:

   `http://www.cs.nyu.edu/~mohri/asr12/spelling_train.txt`
   `http://www.cs.nyu.edu/~mohri/asr12/spelling_test.txt`

   where each line starts with a misspelled word and is followed by a possible correction.

2. create an edit transducer where the edit operations are limited to a single alphabet symbol or two consecutive ones. To further limit the number of possible bigram edits, you could allow for example only transpositions (e.g., *ab* replaced by *ba*).

3. learn the weights of this transducer by using (only) the training set. You could use for this purpose the EM algorithm described in class, as well as other alternative techniques you design.

4. measure the quality of the edit transducer learned by applying it to the test set. To do so, use the edit transducer to return the most likely one or two outputs for each input and measure their edit-distances with respect to the reference. Keep the smallest edit-distance for each input (oracle edit-distance) and average over all the test set.

This is a project and not a homework assignment, so you should explore different solutions and techniques, possibly extend the questions, and not necessarily limit yourself to what was just described.