TRA301/COS401: Homework 5
Due Date: April 23, 2009

1. **Example-based Machine Translation (EBMT):** In this exercise, you will learn to build an example-based machine translation system.

   (a) (10 points): In the file “5-1a.txt”, you are given pairs of English and Spanish sentences; each pair is separated by a empty line. Use these sentences as the example database and translate the Spanish sentences in “5-1a.sp.txt” using an exact-string-match-based EBMT approach. Show the first 10 non-empty translations output by your translation system along with its Spanish sentence.

   (b) (15 points): Evaluate the translated text against the reference translations provided in “5-1a.en.txt”, using string-edit-distance metric. You could adapt the Edit-based FST you created for Homework 2a or write your own code to compute the string-edit-distance between two strings. All edit operation costs are equal to 1. We can then define the accuracy of a translation H against a reference R as: $Acc(H,R) = 100{*}(1 - \frac{numberofedits(H,R)}{lengthofR})$. Compute the per sentence accuracy and use it to compute the average accuracy for the set of sentences in “5-1a.sp.txt”.

   (c) (15 points): Suggest and implement one method to improve your EBMT system. Show the first 10 non-empty translations output by your refined translation system and the improvement in translation accuracy using the metric in 1(b) for the sentences in “5-1a.sp.txt”.

2. **Statistical Machine Translation (SMT):** In this exercise, you will build an IBM-1 alignment model and induce a weighted bilingual lexicon from the alignment.

   (a) (15 points): Using the corpus of English-Spanish sentences from “5-1a.txt”, implement the IBM model 1 algorithm to induce a word-alignment between the words of each sentence pair in this corpus. Show example alignment pairs output by your program for the first 10 sentence pairs in “5-1a.txt”.

   (b) (10 points): Extract a bilingual dictionary from the word-alignment and compute $P(e|s)$ – for each Spanish word (s) the probability of it being translated to an English word (e). For the Spanish words in “5-2b.txt”, show the three highest probability English words and their probabilities.

   (c) (10 points): Using the lexical translation probabilities, design a word-substitution-based translation system that generates English translations (not just one translation) in the order of decreasing translation probability for the Spanish sentences in “5-1a.sp.txt”. The translation probability is the product of the probabilities of the lexical translations for that sentence. Show the five highest probability translations for the first five Spanish sentences in “5-1a.sp.txt”.

   (d) (15 points): Build a bigram English language model from the English sentences in “5-1a.txt”. Rescore the alternate translations in (c) using the language model, by combining the translation model probability and language model probability for each hypothesis.

   (e) (10 points): Using the metric defined in 1(b); evaluate the output of the first best translations produced in 2(c) and 2(d) for all the sentences in “5-1a.sp.txt”, using “5-1a.en.txt” as the reference translation.