# 11-731: Machine Translation

## Homework Assignment #4:

Out: Wednesday, February 7th, 2011

Due: Wednesday, February 14th, 2011

In this homework you will perform the next step of building a Statistical Machine Translation system: extracting phrase translations from word alignment models. You will use the same data you used in Homework 3.

As discussed in the class, to extract phrases we first generate GIZA alignments for both directions of training (Spanish->English, and English->Spanish), then combine the Viterbi paths and extract phrases using different heuristics. In this homework you will investigate several such heuristics.
You will use a script from the Moses[1] package for this purpose. Use the setup provided at /afs/cs.cmu.edu/project/cmt-55/lti/Courses/731/homework/HW4.

Take a look at the script train-phrase-model.perl. It is arranged into several steps.

In this homework, you will proceed up to step 6. Steps 1 and 2 involve generating GIZA alignments that you did in the previous homework. We will repeat it with this script.

It is possible to run only a particular step in the training script (using first-step and last-step parameters).
Here we use the source/target convention as in the lecture slides (i.e. when you translate from Spanish->English, Spanish is the source and English is the target).

Your tasks are:

1) Run steps 1 and 2 in the script and generate GIZA alignments for both directions of training.
```
train-phrase-model.perl --root-dir . --f es --e en --corpus corpus/training --giza-f2e giza.es-en
--giza-e2f giza.en-es --max-phrase-length 7 --first-step 1 --last-step 2
```

2) Generate combined word alignments using UNION heuristic (step 3 in the script). You can specify the heuristic with the alignment parameter.

[1] http://www.statmt.org/moses/

```
train-phrase-model.perl --root-dir . --f es --e en --corpus corpus/training --giza-f2e giza.es-en
--giza-e2f giza.en-es --max-phrase-length 7 --first-step 3 --last-step 3 --alignment union
```

3) Extract phrases, score them and generate the final phrase table (steps 4 to 6 in the script). You can use the same command as in 2) but change first-step and last-step appropriately. When completed, the phrase table will be in model/phrase-table.gz. Take a look at the phrase table (you can use zcat).

4) Analyze the phrase table using the two scripts provided (AnalyzePhraseTable1.sh and AnalysePhraseTable2.sh). Use the analysis to find the following details:
   a. For source phrases up to length 3, give the following statistics:
      i. the number of different source phrases
      ii. average number of translations
   b. Maximum number of translations seen for any source phrase of any length
   c. Give the output of AnalyzePhraseTable2.sh

5) Similar to what you did in homework 2, write a script that will read source phrases of length one from the phrase table and compute the coverage statistics for the test set test/dev2006.es.pp.lc. (i.e. compute the number and percentage of types and tokens that are not found in the phrase table). The test set is already preprocessed.

6) Repeat 2) to 5) for the following heuristics:
   a. INTERSECTION (intersect)
   b. GROW-DIAG-FINAL (grow-diag-final).

7) Compare the statistics you generated in 4) and 5) for each of the 3 heuristics. State your observations.