Homework #1: Statistical Machine Translation

Anoop Sarkar – anoop@cs.sfu.ca

(1) Implement the Church-Gale sentence alignment algorithm as described in the following paper:

> Gale, William A.; Church, Kenneth W. (1993), "A Program for Aligning Sentences in Bilingual Corpora", Computational Linguistics 19 (1): 75102.
> http://aclweb.org/anthology/J/J93/J93-1004.pdf

The source code is actually part of the paper!

The data for testing your aligner will be Chinese-English parallel data with sentence boundaries in each language already detected. There are no word boundaries in Chinese but the Church-Gale algorithm uses characters anyway. The data is available at /cs/natlang-data/champollion-1.2 or from http://champollion.sourceforge.net/

You may want to convert the encoding for the Chinese data from the original GB2312 encoding to UTF8 to help debug your program using `iconv -f GB2312 -t utf8 < input > output`.

Compare your output alignment with the gold alignment for the files in the `eval` directory using the command:

```
diff -y --suppress-common-lines UN19990209_010.align UN19990209_010.gold.align
```

(2) A *paraphrase* of a sentence is an alternative method to render the same or similar information. Use a language model to find the corpus probability of a corpus and its paraphrase. Report which version is *better* according to the language model. The data is available at /cs/natlang-data/kjv-bbe

A 5-gram language model in ARPA format is available at:
/cs/natlang-data/wmt10/lm/eparl_nc_news_2m.en.lm

The kenlm language model package is available at: http://kheafield.com/code/kenlm/. For x86_64 machines the LM in kenlm binary format: eparl_nc_news_2m.en.binlm. Loading the binary version is much faster.

(3) Build a machine translation system using Moses for a language pair of your choice from the European Parliament (EuroParl) corpus: http://statmt.org/europarl. Follow the step by step instructions given in http://www.statmt.org/moses/?n=Moses.Tutorial.