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:


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/](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_news2m.en.lm

The kenlm language model package is available at: [http://kheafield.com/code/kenlm/](http://kheafield.com/code/kenlm/)  
For x86_64 machines the LM in kenlm binary format: `eparl_news2m.en.binlm`. Loading the binary version is much faster.