CA446: Statistical Machine Translation
Lab Exercises

Friday, 6th March 2015

Download and unzip the *Lab2* data from the course website.

1. Part One: Language Modelling

   (a) Write a program which reads in a corpus of sentences and trains a unigram language model on these sentences.

   (b) Using the *toyCorpus.txt* file as your training corpus, calculate the probabilities of the following sentences:

   a a a

   a x a

   (c) Using the *wikiOscars.txt* file as your training corpus, calculate the probabilities of the following sentences:

   Janet won

   Joan won

   (d) Modify your program so that bigram language models can also be trained.

   (e) Modify your program so that it performs add-one smoothing. You can assume that the vocabulary is all the distinct words in the training corpus plus all words in the test sentences.

2. Part Two: Evaluation continued

   (a) Finish your evaluation script from last week's lab. If you are calculating n-gram precision, make sure that you clip the number of times a word in the source can be counted by looking at the number of times it occurs in the reference.

3. Part Three: SMT systems – If you're feeling adventurous, download one of the following systems onto your laptop.

   (a) Moses: `http://www.statmt.org/moses/`

   (b) Joshua: `http://joshua-decoder.org/`

   (c) cdec: `http://cdec-decoder.org/`