CA446: Statistical Machine Translation
Project Ideas
Wednesday, 26th March 2014

1. **Evaluation Script:** Write a wrapper evaluation script that prints out a complete evaluation of a system using BLEU, NIST and METEOR (and perhaps more).

2. **Evaluation:** Do a comparative evaluation of the output of Google Translate, Bing Translate and Online Moses (http://demo.statmt.org/) for various language pairs.

3. **Twanslate:** Train a system that translates tweets.

4. **The effect of the language model in domain adaptation:** If you train an SMT system on one subject (e.g. medical abstracts) and then attempt to translate text on a different subject (e.g. parlimentary pro- ceedings) you will observe a drop in performance compared to when you train and test on the same domain. Examine to what extent this domain effect can be alleviated by training the language model component of your system on text from the target domain.

5. **Irish-English translation:** Train a system that translates from Irish into English and back again. You will need to do some research into what kind of parallel corpora are available for Irish and English. This is a good place to start: http://borel.slu.edu/corpas/ (Kevin Scannell)

6. **Source side re-ordering:** Devise methods to change the order of words in language A into an order which looks more like the sentences in language B. These methods can make use of linguistic information. Apply these methods to the language A sentences in an A-B parallel corpus. Train an SMT system, and see whether this re-ordering improves over a baseline system in which no re-ordering is performed before translation.

7. **Post-editing:** Write a post-processor that corrects the output of a machine translation system using linguistic knowledge

8. **Translation using a pivot language:** Build a translation system that performs an indirect translation from language A to language C. It

does this by first translating from language A into language B and then translating from language B into language C. To train this system you will need to have a parallel corpus for A-B and B-C. To test the system you will also need some A-C sentence pairs. You can compare this approach to one in which you train a system which directly translates from A to C using whatever limited data is available. You can also experiment with different pivot languages (language B).

9. **English-French word alignment**: Using the Hansards dataset:

   `http://www.cse.unt.edu/~rada/wpt/index.html#resources`

   try to improve English-French word alignment.