CA446: Statistical Machine Translation
Lab Exercises

Friday, 13th March 2015

1. Download and unzip the *Lab3Data* from the course website.
2. Login to the student server *student.computing.dcu.ie* using **ssh**.
3. Create a new directory called *CA446*.
4. Create a subdirectory of *CA446* called *Lab3*.
5. Create a subdirectory of *Lab3* called *Data*. Copy the following files from *Lab3Data* to this directory:
   (a) train.fr
   (b) train.en
   (c) testset.fr
   (d) testset.en
6. Create another subdirectory of *Lab3* called *fr-en*.
7. Create a subdirectory of *fr-en* called *corpus* and copy the four files from the *Data* directory into this directory.
8. Create a subdirectory of *fr-en* called *train* and copy the file *train.sh* from *Lab3Data* into this directory. Modify it so that it references your paths.
9. Create a subdirectory of *fr-en* called *test* and copy the file *test.sh* from *Lab3Data* into this directory.

10. Create a subdirectory of *fr-en* called *lm* and copy the file *train-lm.sh* from *Lab3Data* into this directory. Modify it so that it references your paths.

11. Train a language model by running the *train-lm.sh* script from the *lm* directory.

12. Train a translation model by running the *train.sh* script from the *train* directory.

13. Run the decoder by running the *test.sh* script from the *test* folder. When it is finished you should find the translation in the file *testset.train.en*.

14. Get the BLEU score for the *testset.train.en* translation. Hint: Look at last week's worksheet.

15. Create a directory called *en-fr* at the same level as the *fr-en* directory, and repeat the train/decode process in the opposite translation direction.

16. Take a look at some of last year's reports.

17. Download *IWSLT09* data from the link given in `http://www.computing.dcu.ie/~sdandapat/SMT_lab/mt.html`. You can choose either Chinese-English or Turkish-English. Use 1500 parallel sentences to create the training set. Use a disjoint 30 sentences to test the system. Train an MT Engine using this dataset and evaluate the system accuracy.

18. if you still got time please add the following two lines in the train.sh

    ```
    --first-step 1
    --last-step X
    ```

    Note: Try the X from 1 to 9 and check the output of train.log, the directories and files generated.