CA446: Statistical Machine Translation
Lab Exercises

Friday, 27th February 2015

1 Part One: Evaluation

1. Write a program to evaluate a translation (sentence) against a reference translation (sentence). The input to your program should be two sentences, the first representing the translation to be scored and the second representing the gold standard reference translation. The output of your program should be a score. You can use any programming language you like.

2 Part Two (Using an SMT System)

1. Login to the student server *student.computing.dcu.ie* using **ssh**
2. Create a new directory called *CA446*.
3. Create a subdirectory of *CA446* called *Lab1*.
4. Create a subdirectory of *Lab1* called *Data*. Download the *Lab1Data.zip* file from the course website and unzip.
5. Create another subdirectory of *Lab1* called *Fr-En*.
6. Create a subdirectory of *Fr-En* called *orig* and copy the files from the *Data* directory into this directory. The directory should contain the following files:

- *train.fr*
- *train.en*
- *testset.fr*
- *testset.en*

7. Run the following commands from your *Fr-En* directory:

   (a) Language Model – make a directory for your language model

   ```
   cd ~/CA446/Lab1/Fr-En
   mkdir lm
   ```

   Now train a language model

   ```
   MOSESPATH=/usr/local/mosesdecoder

   ${MOSESPATH}/bin/lmplz -o 5 -S 80% \
   -T /tmp < ./orig/train.en > ./lm/en.arpa
   ```

   Then binarize the language model

   ```
   ${MOSESPATH}/bin/build_binary \
   lm/en.arpa lm/en.blm
   ```

   (b) Translation Model – now you must create the translation model

   ```
   cd ~/CA446/Lab1/Fr-En
   mkdir work
   cd work

   MOSESPATH=/usr/local/mosesdecoder
   USERNAME=student_username

   ${MOSESPATH}/scripts/training/train-model.perl \
   -root-dir train -corpus ../orig/train \
   -f fr -e en \
   -alignment grow-diag-final-and \
   -reordering msd-bidirectional-fe \
   -lm 0:5:/users/${USERNAME}/CA446/Lab1/Fr-En/lm/en.blm:8 \
   ```

```
-external-bin-dir /usr/local/bin/ \
>& training.out &
```

(c) Decoding – produce the output translation for the testset you have in your *orig* directory

```
cd ~/CA446/Lab1/Fr-En
mkdir test
cd test

MOSESPATH=/usr/local/mosesdecoder

${MOSESPATH}/bin/moses \
-f ../train/model/moses.ini \
< ../orig/testset.fr > trans.en
```

(d) Evaluation – evaluate the system accuracy in terms of BLEU

```
cd ~/CA446/Lab1/Fr-En/test
MOSESPATH=/usr/local/mosesdecoder

${MOSESPATH}/scripts/generic/multi-bleu.perl \
-lc ../orig/testset.en < trans.en
```

8. Translate the French test set (in the *orig* directory) into English using Google Translate. Save the translation as **testsetGoogle.en** in your *test* directory and evaluate it as follows:

```
MOSESPATH=/usr/local/mosesdecoder
cd ~/CA446/Lab1/Fr-En/test

${MOSESPATH}/scripts/generic/multi-bleu.perl \
-lc ../orig/testset.en < testsetGoogle.en
```

Then, manually tokenize and lower case the **testsetGoogle.en**, save it, run the above command again.

Compare the BLEU score achieved by the *Fr-En* system you trained to the BLEU score achieved by Google Translate. Manually compare the actual translations, i.e. look at them!

9. Create a subdirectory of *Lab1* called *En-Fr* and then train a system to translate from English to French. You will first need to create an *orig* subdirectory that contains the same data as *orig* in *Fr-En*.