11-731: Machine Translation

Homework Assignment #2:

Out: Monday, January 24, 2011

Due: Monday, January 31, 2011 (email thuylinh@cs.cmu.edu before class)

In the course of the semester you will build a Spanish-English translation system. In this homework assignment you will start with preparing and analyzing the data. This is the data, which has been used in recent ACL Workshop on Statistical Machine Translation (http://www.statmt.org/wmt08/). Please, use the copy of the Spanish-English data provided at

`/afs/cs.cmu.edu/project/cmt-55/lti/Courses/731/homework/HW2/` . You will find the training corpus

`europarl-v3b.es-en.[es|en]`

and the (development) test corpus

`dev2006.[es|en]`

The files are sentence aligned. i.e. n^{th} sentence in the English corpus is the translation of the n^{th} sentence in the Spanish corpus.

Your tasks are:

i. Tokenization: Notice that the training data is still in raw format. Write a simple program to separate the punctuations from words. Are there any special cases, which might need special treatment? Give one or two sentences for both languages showing the differences due to tokenization.

ii. Corpus statistics: Give corpus size (number of sentences and word tokens) and vocabulary size (word types) for training and development data, for the original data and the processed data. What is the average sentence length? How long is the longest sentence? Compare Spanish-English, and original-tokenized data.

iii. Sorted word frequency list: Write a simple script to count how often each word occurs in the corpus. Sort the list according to frequency. Do this for Spanish and English. What is the percentage of singletons, i.e. word types, which appear only once in the training corpus? Look at the high frequency words in the lists. Do you see any interesting similarities or differences?

iv. Vocabulary growth: We want to observe the growth of the vocabulary size as the corpus size increases. Write a simple script to compute the size of the vocabulary at different levels of the corpus. Select the top 1k, 2k, 5k, 10k, … 200k sentences, up to full corpus. Using your script from step (i) you can get vocabulary size and number of singletons. Plot the data and state your

observations. Ideally, your plot shows four curves: vocabulary size ES and EN, and singletons ES and EN. You don't need to submit your program.

v. Vocabulary analysis: In each corpus, how many entries contain digits [0-9], how many entries contain punctuation, how many entries contain a mix of characters [A-Za-z] and digits [0-9] ?

vi. Unbalanced sentences: For each Spanish and English sentence pair, compute the length ratio (i.e. # Spanish words/# English words). Look at sentence pairs with ratio >= 2.0. Select one example, where the sentences are not a good translation pair. Why do you think it is not good? Give one example, where the translation is correct, despite the unbalanced number of words.

vii. Coverage: For the given Spanish development test set, what is the percentage of words (types and tokens) that is not found in the training data (unseen words)? Write a simple script, which takes a vocabulary (or the word frequency list from above) and a file, and calculates type and token coverage. Run this over the different corpora generated in (iv) to see how coverage improves with corpus size. List the number of unseen words (types and tokens) and the out-of-vocabulary rate.

viii. What other processing methods would be helpful in reducing the vocabulary size? Why do you think that would be helpful for machine translation?

Provide links to your programs/scripts and the processed test corpus.