COMS 6998-4, Fall 2012
Statistical NLP for the Web

Homework 3

Due: Nov. 30 (Friday), 2012 (11:59pm)

Total Points: 100

Machine Translation, Language Model and MapReduce

In this homework, you will learn how to build a simple Machine Translation model and use the trained model to decode foreign language sentences. You will also learn how to build language models in MapReduce framework.

Q1. EM Algorithm for Model 1 Alignments

Data 1 : http://www.statmt.org/mtm2/data/fr-en.tiny.tgz
Data 2 : http://www.statmt.org/europarl.tgz

(1) Pick your favorite foreign language. Pick a few short sentences of varying length in English and translate them into your favorite foreign language. You can also find such translation pairs from Data 1 (shorter sentences) or Data 2 (slightly longer sentences). [5]
(2) Write EM algorithm described in the class to build word translation table. [30]
(3) Translate a few English sentences from your corpus into your target foreign language using simple word translation model built in Q1.2. Report automatic translation of random 5 English sentences. [10].
(4) Describe the potential problems that may arise (if any) when you use your algorithm in Q1.2 for very long sentence pairs and how you may address them [5].

Q2. Language Model in MapReduce

Data 3: /home/smaskey/CS6998-0412/hw3/lm_data

(1) Implement a simple n-gram counter that counts number of unigrams, bigrams and trigrams in a subset of data files provided in Data 3. [7]
(2) Implement a MapReduce version of your counter and run your implementation in Amazon Hadoop. Try different number of nodes in your Hadoop job and report the change in running times. [30]
(3) Create a bigram language model using statistics found in Q2.2 [13]

Extra Credit :
Use the language model you built in Q2.3 to rescore the translations in Q1.3. Do translations improve?