# Homework 2

Sivakarthik Gade

## I. WRITTEN QUESTION

Explain in your own words how language models are used in the following tasks.

a) **Machine Translation:** In this field, language models are used for statistical machine translations. Prior to this, rule based and example based approaches were being used. With the introduction of statistical machine translation (which uses language models), machine translation has revived and by far this is the most widely used machine translation technique. Language models are used in this case to get the probability distributions about a string in the target language to be the translation of the string in the source language.

b) **Information Retrieval:** In this field, language model is usually associated with each document. A language model is trained for each document corpus, and is used in finding the document(s) that best match given query i.e. The language model which gives the best likelihood for the given query is chosen. Usually unigram language models are used for this purpose, as they tend to give faster response than other n-gram counterparts.

## II. PROGRAMMING PART

### Instructions to compile and run the code:

1. Implemented the assignment in Java. Need Oracle/Sun JDK setup on the target machine to compile and run the code. (Used JDK 1.7.0_45 during development)
2. Extract all java files into some location.
3. Compile instruction through command prompt: ‘javac *.java’
4. Run instruction through command prompt:
   a. java bigramtrain -text hw2_train.txt –lm lm.txt
   b. java bigramtest –text hw2_test.txt –lm lm.txt

Perplexity value generated with bigrams on the given test set using the learning model based on given training set: 58.31670706998582

Observations:

1. Have noticed perplexity value is inversely proportional to the training data size, keeping all other factors constant. As training set size increased, the perplexity tended to decrease. This can be attributed to the fact that as more training data is seen, the language model learnt tends closer to the actual probability distribution. As a result, the accuracy of prediction tends to increase (and perplexity decrease).

2. The perplexity tended to higher value in case of unigram model than bigram model. This is supporting the existing evidence that as ‘n’ increases, respective n-gram models tend to be more accurate.