Information Retrieval (CS F469)

Assignment – 2

Cross Lingual Document Translator

Documentation

A Sai Rahul 2013B1A7687P

Akshith Reddy S 2013B3A7673P

Sai Teja K 2013A7PS034P

Akhil Reddy P 2013A7PS152P

Objectives:

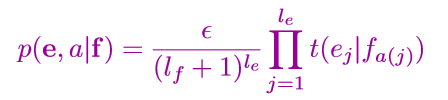
* To implement a Cross Lingual Document Translator based on statistical machine translation model.
* To compare the generated translation against an accurate translation using Cosine Similarity and Jaccard Coefficient.

Introduction:

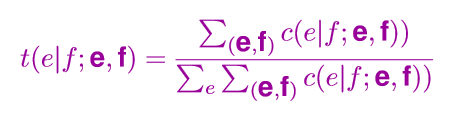
The cross language document translator is based on the word to word statistical machine translation model. Statistical machine translation model relies on translation examples contained in a parallel corpus. The statistical machine translation model has been implemented using the IBM Model 1 and Expectation Maximization algorithm.

IBM Model 1 and EM Algorithm:

* IBM Model 1 is a word alignment model, which uses an Expectation Maximization Algorithm to compute the lexical translation probabilities in parallel texts.
* Translation probability in IBM Model 1 for a foreign sentence f length lf to an english sentence e of length le with an alignment of each English word ej to a foreign word fi according to the alignment function a : j → i



* After assigning probabilities and estimating the count until convergence using the EM algorithm, we can estimate the model



Implementation:

1. Pre-Processing
   * In the initial step the parallel corpus is parsed.
   * Special characters are removed from the parsed text.
   * The parsed text is then subjected to case folding and tokenization.
   * The terms generated are stored in a nested dictionary.
2. Statistical Machine Translation
   * Initial probabilities are assigned and counts (weighted by probabilities) are collected to estimate the model after 10 iterations, using a multi-threaded approach to avoid overheads.
   * The translator accepts either an English or a French text documents and generates corresponding translated documents in French or English respectively.
3. Similarity Computation
   * Cosine similarity and Jaccard coefficient are used to calculate the similarity scores between the generated translation and an accurate translation.
   * Cosine Similarity is a similarity measure between two non-zero vectors that measures the cosine of the angle between the vectors. The generated translation and an accurate translation are represented as vectors using the tf-idf weighting.
   * The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. The generated translation and an accurate translation are represented by sets that contain all the terms contained in the respective document.

Github Repo : http://github.com/TEJATJ/clir