**CROSS LANGUAGE DOCUMENT TRANSLATOR (CLDT)**

**Group 17**

Aaron Fernandes – 2019H1030512P

Sri Krishna Manoj – 2019H1030020P

Jai Singhal – 2019H1030021P

Aswin B – 2019H1030022P

Gaurang Bansal – 2014HS030128P

Gaurav Kumavat – 2016A2PS0737P

**Introduction**:

Cross Language Document Translator(CLDT) is a tool to translate documents between two languages. It makes use of Cross Language Information Retrieval(CLIR) to perform the translation. CLIR is a subfield of Information Retrieval that deals with retrieving information independent of the language used.

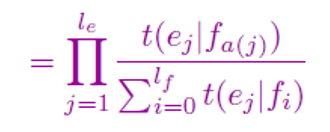
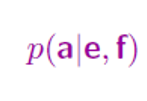
**IBM Model - 1:**

IBM Model – 1 is a statistical machine translation model, which is used to perform lexical translation. It does not consider rearrangement, or dropping and adding the words.

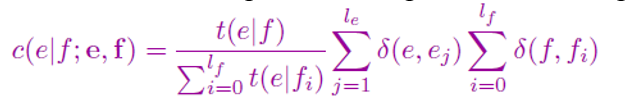
There are two main steps in this model. They are:

1. Expectation Step
2. Maximization Step

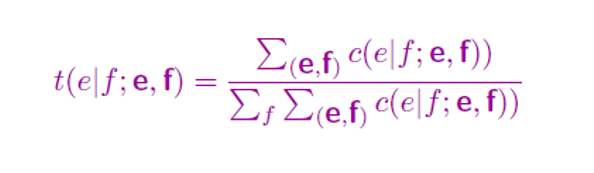
In Expectation Step, we assign probabilities to word pairs based on the number of alignments that they occur in. In Maximization step, we take the count of each of these alignments and normalize them to generate word-pair probabilities. These two steps are performed iteratively till the probabilities converge.



Probability of an alignment



Number of times the word ‘e’ in English is mapped to a foreign word ‘f’



The probability of translating English word ‘e’ from foreign word ‘f’

**Code:**

There are two files an

* Ir\_project\_notebook.ipynb : for training and testing purpose
* code.py : for testing only

**Important function in notebook:**

1. **Training part:**
2. ***cleaningInputFiles():***

It takes the dataset as input and removes non alphabetic characters and builds a new dataset for both the files.

1. ***ibmModel1Train(english, dutch, transition\_prob, iterations)***

It takes parameter as follows

English: English dataset

Dutch: Dutch dataset

Transition\_prob: Probabitiy Dictionary

Iterations: Number of iterations

It creates all the pickle file of the translation with their respective probability.

  This function is responsible for taking the test document in both language as input parallel corpus. As the no of iterations and the amount of lines is specified in the code it continuously calculates the word pair probability and stores them in a pickle file after each iteration. In this way a pickle file is created for each and every iteration of model1. Before writing onto a file we use a Dictionary data structure to temporarily store the word pair probabilities.

1. ***ibmModel2Train(english, dutch, transition\_prob, iterations)***

It takes parameter as follows

English: English dataset

Dutch: Dutch dataset

Transition\_prob: Probabitiy Dictionary

Iterations: Number of iterations

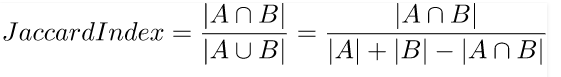
The final pickle file generated by model1 is taken as input to model2 and model2 starts training on the same document with prior probabilities equivalent to the ones generated by model1’s last iteration.

1. **Testing part:**
2. ***get\_cosine\_similarity():*** It calculates and returns the cosine angle similarity between 2 documents present in the higher dimensional space.



1. ***jaccard\_coefficient():***

 It calculates and returns the jaccard coefficient, a measure of similarity between any 2 documents given as input to the function. The list is typecasted to a set for easy calculation of union and intersection.



1. ***getTranslations()***

Retrieves the translation of all words with maximum probability and returns the same.

1. ***run()***

Driver Code for testing.

**Assumptions:**

**Results:**

We use the following parameters to calculate accuracy

1. **Jaccard’s coefficient:**

We are using the jaccard’s coefficient to find the similarity between the resulting translated document to the source document. It is defined as the size of the intersection of the sample set divided by the size of the union of the sample sets:

**2) Cosine similarity:**

Cosine similarity is a measure of similarity between two non-zero vectors that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any angle in the interval (0,π] radians. This results in a judgment that takes orientation and not magnitude into account. We use this similarity index to find the accuracy of our translation along with jaccard coefficient.