NLA Project - Write Up
Cross Lingual Information Retrieval

Submitted By
Aniruddha Deshpande (20161058)
Samyak Agrawal (20161180)

Project Mentor : Nikhil Pattisapu
Faculty Mentor : Dr. Vasudev Varma

Crosslingual information retrieval (CLIR) is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query. This is opposite to that of **Monolingual information retrieval**, where both the query and the output are in the same language. CLIR can be both **Dictionary based** or based on a **Parallel Corpora** of sentences.

Following are the major tasks involved in the process of CLIR :

- **Crawling :** Documents from web are fetched and stored. In our case, the process of CLIR is *Domain Based*, hence the documents to be crawled for different languages should be specific to the chosen *Domain.*
- **Ranking :** The systems produces a list of documents, ranked according to their *relevance to the query*. Several different types of algorithms can be used for this process.
- **Query Boosting :** This task involves boosting the weights of certain important words in a query. This can be done to avoid and reduce the number of output documents that are relevant to only some certain words but not to the overall query. *For eg. A query regarding the Russian Prime Minister shouldn't generate output documents regarding Prime Minister of India. Here the weight of the word 'Russian' is boosted to avoid such a scenario.* What values are to be boosted are learnt with data.

Following are the two approaches that can be chosen :

- **Query Translation :** Here, the query is translated into the target output language and then the search is carried out. This is a relatively easier task as queries are shorter in size.
- **Document Translation :** Here, the search takes place in the given source language and the document to be displayed is translated into the target language of user's choice. It is a difficult task as translation process becomes difficult as the size of the input increases. Bigger the document, harder it is to translate.

Preliminary Tasks Assigned :

1. Choose a pair of favourable language.
2. Choose a single domain to work on.
3. Choose a Document Ranking Algorithm.
4. Crawl and collect documents based on the chosen domain.
5. Implement a simple CLIR system based on the above chosen factors (without Query Boosting)
6. The code must be a modular code. (This makes the task of adding Boosting in the later versions easier.)

Suggested References :

- **TDIL Sandhan :** https://tdil-dc.in/index.php?option=com_vertical&parentid=86&lang=en