# Designing Cross-Language Information Retrieval System using various Techniques of Query Expansion and Indexing for Improved Performance

**Aditi Agrawal[1], Dr. A. J. Agrawal[2]**

*[1]Student M.Tech, CSE Department,
Shri Ramdeobaba College of Engineering and Management, Nagpur, India.
[2] Professor, CSE Department,
Shri Ramdeobaba College of Engineering and Management, Nagpur, India*

---***---

**Abstract -** *Classical Information Retrieval is finding out of the documents most relevant to a user's query, from a large store of documents. A search engine performs IR by retrieving relevant web documents from the internet. Cross Language Information Retrieval permits the user to retrieve their documents in other language then the query. Some systems use resources such as bilingual dictionaries to translate the user's original query and other systems use machine translation to translate the documents. Problems arise due to ambiguity in language. In this paper we will discuss previous work in CLIR. Four papers are considered for literature review and the proposed approach is discussed*

***Key Words*: Cross Language Information Retrieval, Hindi-English Dictionary, Disambiguation.**

## 1.INTRODUCTION

With the explosion of knowledge on the web, it became necessary to break the obstacle of language for the IR systems, CLIR is filling the gap of linguistic barrier by allowing a user to search document in different language then the query. Indian languages have gained importance in evaluation recently; User wants information in his native language. In India, about 70% of people know Hindi as a primary language while based on a recent human development survey; there are only 10.35 % peoples in India who are the English speakers. The main aim to develop a new approach and looking into the limitations of existing approaches is to find out all the relevant information from CLIR with higher recall and with no or very less amount of irrelevant information retrieved by the query provided by the user. System performance of CLIR is still poor as compared to Monolingual performance.

## 2. LITERATURE REVIEW

The first paper is "Query Expansion for Cross Language Information Retrieval Improvement" written by Benoit Gaillard, Jean-Leon Bouraoui, Emilie Guimier de Neef and Malek Boualem in year 2010.The proposal of this paper is to overcome the problem of differences between translated data and human language that is induced by translation software by using Query Expansion (QE). QE consists of adding new words to the initial query. Terms that are contained in the initial query are matched with the documents.

CLIR module: The CLIR module adds a translation service to the IR engine. The model on which this paper is based involves translating the contents before indexing them. The original text is kept in the memory of the engine; only the translation is indexed in order to present the original version of texts to users.

QE module: It is based on the TiLT platform4. Its aim is to provide, for query terms, some corresponding expansion terms. It can perform QE according to five modes. The system performs the tasks searching through a document and indexing it. It accepts simple but also complex queries. Indexing and search engine are the backbone of the architecture, connecting the CLIR and QE modules. The collection of documents is searched using search engine. The engine searches the index of translated metadata with the expanded query after the QE module expands the initial query.

Second review paper is "Enhanced Query Expansion in English-Arabic CLIR" written by Abdelghani Bellaachia and Ghita Amor-Tijani in Year 2008. The paper tells that Query Expansion proved to be effective. Using the top retrieved documents the concept of a query is enhanced by adding to its context related terms. Those documents are retrieved using the initial query. Related terms are extracted from the presumed relevant documents using co-occurrence analysis, based on the assumption that if two terms co-occur then they tend to be related. Then the final expanded query is formed by adding those terms. The optimum effectiveness could be reached by applying Disambiguation. Using the DQE technique, The approach of QE enhanced by a thesaurus-based disambiguation, it was tried on the basis of consideration that not all expanded terms are necessarily related to the query. The co-occurring words with the query terms in the set of top ranked documents are added to the translated query using QE. Using DQE, the terms that are directly related to the initial query are considered in the expanded query and analyzed. After the query is tokenized and stemmed, query terms are translated. The first approach is, to extract spelling variants of the transliterated word the Transliteration N-Gram (TNG) technique is applied. Using this approach, a set of possible transliterations is obtained by using the ngram approximate string matching technique on a transliteration and the stemmed terms in the index of the document collection. The other approach (eTNG), referred to as the enhanced TNG, Transliterations were extracted using TNG and POS disambiguation was applied. Query expansion is performed using context-related terms from the top n documents. Further disambiguation of those expanded terms is done by applying WordNet disambiguation. A final retrieval was carried out.

Different runs were carried and compared:

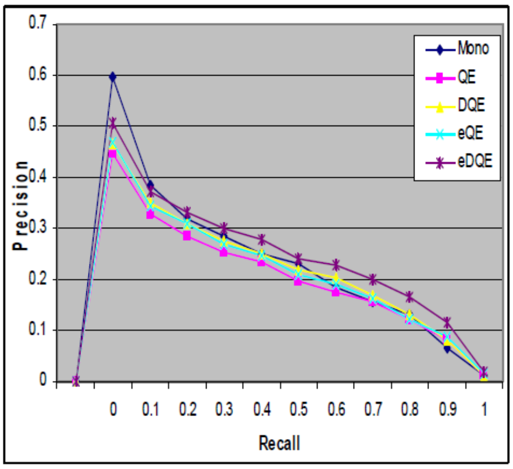

**Fig-1:** Precision-Recall Graph of the Mono, DQE, QE, eQE, and eDQE Runs.

(Figure Reference: - “Enhanced Query Expansion in English-Arabic CLIR” written by Abdelghani Bellaachia and Ghita Amor-Tijani in Year 2008.)

eQE: Translated queries processed with the eTNG technique were treated with Query expansion

DQE: Disambiguation using WordNet was applied on expanded queries.

eDQE: expanded queries processed with eQE were treated with Disambiguation using WordNet

Mono: Arabic queries were used to search the Arabic document set. The monolingual run was used to evaluate the performance.

DQE technique. Al-Stem was used to stem both the Arabic document and the Arabic query set.

QE: Translated queries processed with the TNG technique were treated with Query expansion

The third review paper is “Hindi - English Based Cross Language Information Retrieval System for Allahabad Museum” Authors are Vivek Pemawat, Abhinav Saund and Anupam Agrawal written in Year 2010,The paper talks about the system developed for CLIR. The documents and images for query processing were related to Allahabad museum. The languages used were Hindi and English. The documents were stored in English language. The user has a choice to input a query keyword in either Hindi or English. The system retrieves the relevant documents and displays them in the desired language. Users can get the relevant information from the documents which are in free text format. The steps followed are

A. Algorithm for Retrieval System

Conversion of Hindi words to English words: For the conversion of Hindi words to English words, a dictionary database was used and for those not in the dictionary database, Hindi character to English character mapping has been used. So in this way, map data structure is used and conversion is done.

Retrieval of documents and images: Vector based model was used for the retrieval of documents and images, all the documents were in English language in free text format. Preprocessing of documents is done by stemming and stop word removal. After that each documents is represented by vector in vector space.

Clustering, it is a method to process the data and classify documents into different classes and group so that we can know what object lies in which category and can find the similarity easily. By applying clustering in vector space complexity of search time was reduced. In this paper they have discussed the importance of spell checkers which is nowadays a special tool for millions of users. They have worked with web as corpus in order to design flexible and dynamic spell checker. Edit distance algorithm has been used

B. Conversion of English Documents to Hindi Documents

Finally after retrieval of documents from the above retrieval algorithm they showed the output in the language of user choice. If it is English then there is no need to convert but if it is Hindi then conversion is required. For that Google API was used. With the use of driver manager class connection was maintained with the given database url.

Fourth paper is “Cross Language Information Retrieval: In Indian Language Perspective” written by Authors Pratibha Bajpai and Parul Verma in Year 2014. Through this paper review of the work done in the field of Indian languages by various researchers for CLIR system is provided. Analysis of CLIR approaches for Indian languages: Table analyzes the performance of various methods used by the researchers for Indian languages.

## 3. PROBLEM STATEMENT

### 3.1 Discussion

Nowadays Cross Language Information Retrieval has become a crucial part. User wants to read the documents in language which they understand best. For that reason we need to develop this approach to provide accurate and better results. In particular for Hindi language not much work has been done and there are many limitations with the proposed approaches.

### 3.2 Problems

The existing CLIR system for English Hindi has some of the following limitations

- Translation Disambiguation
- Out-of-Vocabulary problem
- Translating phrases
- Named entities

In our system we will try to develop a system for CLIR that would overcome the existing limitations and provide better results.

## 4. PROPOSED APPROACH

The project is divided in two parts First Processing the Query and second Building the Index. This system will include a dictionary based on particular domain. The first module consists of index which will contain the English words and their corresponding Hindi words and relationships. We will map the index in a way that the documents to be retrieved are accurate. The documents will be processed and stored in index. The second Module is processing the query in which query processing will be done to simplify the query the query can be in both English and Hindi we will keep the count of Hindi and English words so during retrieval of document more weight age can be given to the documents in the language with higher score. All the terms will be stored in English as well as Hindi. The terms id will be allotted. The English term will also contain the term id of its respective Hindi term and the documents that contain that English term and vice versa. So when user types query in English and wants answer in Hindi the Hindi term id will be selected and the documents in Hindi corresponding to that term will be retrieved.

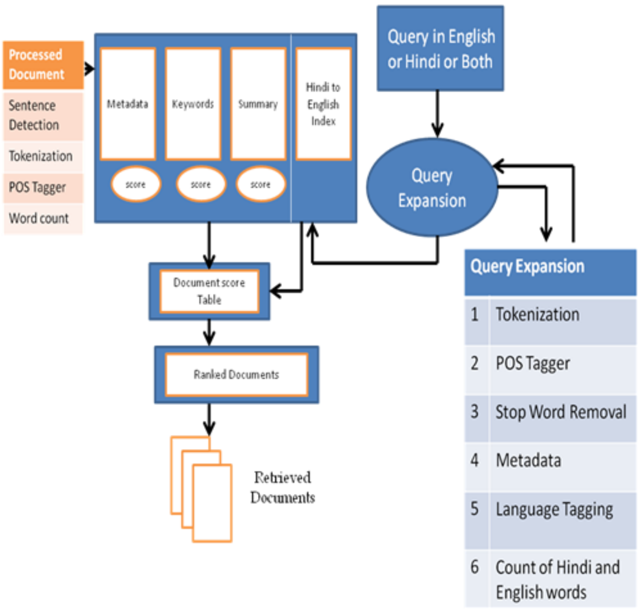

**Fig-2:** Proposed Architecture

## 5. CONCLUSION

For the input query for a particular domain in Hindi or English the documents will be retrieved. The documents can be both in English and Hindi. The score for individual Index will be calculated using tf-idf and total score will be stored in Document score table. The Processed query will be given as input to the index where the Hindi to English index is available the document ids will be obtained and documents can be retrieved from score table depending on score the Hindi English word count will be used for ranking documents. The Precision and Recall will be used as accuracy measures to evaluate the system.

## 6. REFERENCES

[1] Pratibha Bajpai, Parul Verma "Cross Language Information Retrieval: In Indian Language Perspective" International Journal of Research in Engineering and Technology (IJRET) Jun-2014.

[2] Benoit Gaillard, Jean-Leon Bouraoui, Emilie Guimier de Neef, Malek Boualem "Query Expansion for Cross Language Information Retrieval Improvement" 2010 IEEE

[3] Vivek Pemawat, Abhinav Saund, Anupam Agrawal "Hindi - English Based Cross Language Information Retrieval System for Allahabad Museum" 2010 International Conference on Signal and Image Processing

[4] Abdelghani Bellaachia and Ghita Amor-Tijani "Enhanced Query Expansion in English-Arabic CLIR" 19th international conference of database and expert system application.