International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 155-157
e-ISSN:2278-621X

A STUDY AND ANALYSIS ON CROSS LANGUAGE INFORMATION RETRIEVAL

Ruban S[1], Lesleeta Lobo[2] & Shriraksha N Shetty[3]

Abstract: Information retrieval deals with retrieving of relevant information from databases with the use of computerized system.Information is organized as a collection of documents which are unstructured.It locates relevant information on the basis of user's input such as keywords or example documents.Web search engine is an example for Information Retrieval(IR).Cross Language Information Retrieval(CLIR) is a subfield of information retrieval(IR) that deals with retrieving of information written in one language different from the language of the user's query.It is also called as Multi-lingual IR,Cross-lingual IR or Trans-lingual IR.Forexample,usingHindi queries to retrieve English documents.There is a necessity of some mechanisms that retrieves information in a database.The information stored need not be in one language.Then the simple way to search for the information is to scan every item in a database and when the need to translate the language being used arises,there will be need of developing Cross Language Information Retrieval(CLIR).On the web,we are having documents in different languages,multi-lingual documents,images with caption in different languages.A single query should retrieve all such resources.

Keywords: Information Retrieval, Cross language Information Retrieval, Machine Translation.

1. INTRODUCTION

Information Retrieval (IR) system concerns with retrieving of relevant information from repositions like web or large collection of digital data.It is basically concerned with facilitating the user's access to large amounts of information or finding documents of an unstructured text that satisfies a specified information need from within the storage. With an increasingly globalized economy, the idea of finding information in other languages seems to be a mandatory task.

Information-Retrieval [IR] is the act of database storing, searching and retrieving process of information that matches a user's request .Internet is no longer monolingual as the contents in other languages like Hindi in our native place India is growing rapidly. The diversity of languages is becoming barrier to understand in this digital era. So the information-retrieval (IR) has been an essential area of research in these days. It has been found that when user gets the services in local languages, it has been majorly accepted and used by the users. With the explosion of knowledge on the web, it became necessary to break the language barriers for the monolingual IR systems. This may allow the users of IR systems to give query in one language and retrieve documents in different languages. Web search engines are the most visible IR applications. An information retrieval process begins when a user enters a query into system.Cross Language Information Retrieval (CLIR) is a subfield of Information Retrieval (IR) that deals with some querying issues that are not generally addressed by database systems. In Cross Language Information Retrieval (CLIR) user submits a query in one language to retrieve document in another language. Information Retrieval(IR) deals with representation,storage, retrieval and access of a monolingual document collection, whereas CLIR deals with solving the problem of mapping the query that is in other language/s.CLIR system is a system in which a user is not restricted to only one language, it allows to formulate query in one language and then system returns the documents in the other language. In CLIR the language of query and the documents both can be translated .This paper aims at providing an introductory overview to some approaches that deals with CLIR's cross-language facet.This paper also reviews literature on dictionary-based cross-language information retrieval. Dictionary-Based Translation has become increasingly available and isoftently used in the translation of CLIR engines. A dictionary-based approach for the translation is very easy but it has limitations such as ambiguity and lack of coverage.

2. CLIR MODEL

Cross-Language Information Retrieval (CLIR) allows the users to search and read documents in the language that is different from the language of the search terms. Potential users for CLIR are users who find it difficult to analyse a query in their non-native language and users who are multilingual and want to save time by entering a query in one language instead of entering the query in all languages in which the user wants to. Based on translation resources, Cross Language Information Retrieval(CLIR) has been classified into three primary translations i.e. Dictionary-Based translation, Corpora-Based translation and Machine translation. The following diagram illustrates the CLIR Model

[1] Asst.Professor, Department of IT, St.AloysiusCollege(AIMIT), Mangalore.
[2] MSc Software Technology, St.AloysiusCollege(AIMIT),Mangalore
[3] MSc Software Technology, St.AloysiusCollege(AIMIT),Mangalore.

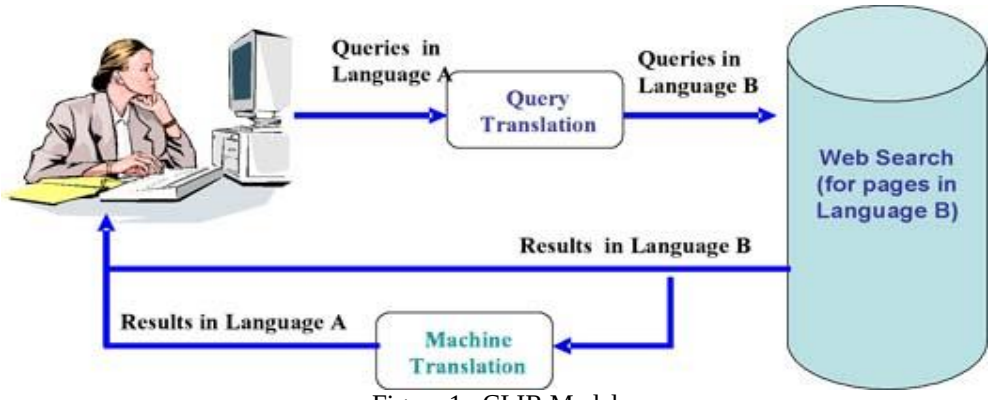

Figure 1 : CLIR Model

Cross-Language Information Retrieval (CLIR) allows the users to search and read documents in the language that is different from the language of the search terms. Potential users for CLIR are users who find it difficult to analyse a query in their non-native language and users who are multilingual and want to save time by entering a query in one language instead of entering the query in all languages in which the user wants to.Based on translation resources, Cross Language Information Retrieval(CLIR) has been classified into three primary translations i.e.Dictionary-Based translation,Corpora-Based translation and Machine translation.

Dictionary-Based Translation has become increasingly available and is often used in the translation modules of CLIR engines. A dictionary-based approach for the translation is very easy but it is having two limitations such as ambiguity and lack of coverage . In Dictionary-Based query translation the keywords are translated to the target language using Machine Readable Dictionaries (MRD). MRDs are electronic versions ofprinted dictionaries, general dictionaries or specific domain dictionaries.

Corpus-based Translation is a document written in one language together with its translation in another language. Large collections of parallel texts are referred to as parallel corpora. Parallel corpora can be acquired from a variety of sources. A corpus is a repository of a collection of natural language material, such as text, paragraphs, and sentences from one or many languages.

Machine Translation is the automatic translation of the text from one natural language to another.Machine translation is a technique that makes use of software that translates text from one language to another language. Machine Translation can be implemented in two different ways. The first way is to translate one language document in the corpora into the language of the user's query. In the second method, the users query in the source language is translated into the target language i.e. in database.

3. EARLIER WORKS

The only area in Information Retrieval which has got many exciting advancements in it, is Cross Language Information Retrieval(CLIR).The goal is to allow all the users to make queries in one language to another language. Obtained documents can then be translated into a language used for the query to allow the user to get the retrieved information. For example, a user makes a query in English, and receives documents back in Hindi. Cross lingual information retrieval for foreign languages like English, French, Chinese etc. has been an appealing area for many researchers from a very long time. But Indian languages have grabbed attention only a few years back. The result obtained by researchers gives mixed results in terms of improvement over monolingual retrieval in Indian language perspective. AnuranSeetha& S. Das (2010) performed translation on Fire 2010 Hindi test collection using Shabdanjali dictionary & query expansion by Hindi Word net. The method proved to be ineffective. This was because, general dictionaries have a lower coverage problem.Inorder to remove this inefficiency Larkey and Connell (2003) used probabilistic dictionary, from parallel corpus for English to Hindi translation and achieved effective cross lingual retrieval. Pattabhi R.K Rao and Sobha. L. (2010) found encouraging results by incorporating Bilingual dictionary and ontology.

4. DICTIONARY-BASED QUERY TRANSLATION

CLIR provides new technique for searching documents through different type of languages across the world .By using the different type of translation techniques CLIR makes it possible to provide the search result in the other language to the language of query. So it will be beneficial for multilingual population regions.

A Dictionary-Based translation is simple when compared to other types of Cross Language Information Retrieval (CLIR), but has weaknesses i.e., ambiguity and lack of coverage. The terms in the query are mutually dependent. Thus, for each of these terms, we need to select the translation that is most consistent with the translations of the remaining terms. The following diagram shows how the dictionary based translation works.

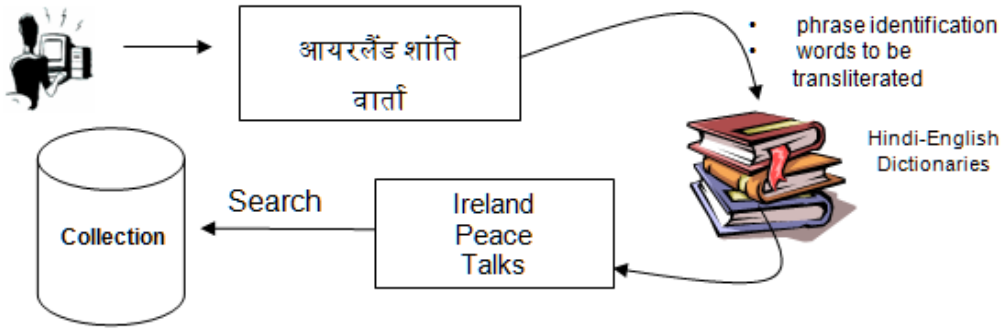

Figure 2 : Dictionary based Translation

5. APPLICATIONS OF CLIR

Some major applications of CLIR are :

1.This CLIR System is helpful for immigration department in airports. For ex. Immigration department interacts with thousands of the Indian native Language people who are not able to understand English Language.

2.Itcan also be used for multilingual population regions so that the people having different native languages can retrieve documents in their native languages.

3. This system is applicable in intelligence departments.

4. The CLIR helps the students in their research work regarding historical places.

Dictionary Based Approach	**Parallel Corpora Based Approach**
In Dictionary Based approach the queries are translated using bilingual dictionaries in which we look for some or all of the translated terms.	In Parallel Corpora Based approach we use a Corpus.
In dictionary based translation the basic approach is to translate word-byword	In Parallel Corpora Based translation query does not need to be translated, since a source language query is matched against the source language component of the corpus,
There can be out of Vocabulary problem, caused by missing query terms in dictionaries	Corpora based methods suffers lack of resource s.

From the above table we come down to a conclusion that Dictionary Based approach is more efficient compared to Parallel Corpora Based approach.

6. CONCLUSION:

The respective work with regard to Indian languages has gained impetus in last decade and there is much to be explored in this field. It is quite obvious from the observations that there is still a scope of improvement in the performance level of CLIR. We presume that the proposed prototype system will prove to be competent with other existing systems. The experimental results show that the proposed approach gives equal/better performance of English-Hindi CLIR system compared to monolingual performance and also helps in overcoming existing problems and outperforms the existing English-Hindi CLIR system in terms of average precision.It is more convenient to translate only the query than the whole documents. Document translation which uses machine translation is computationally expensive and the size of document collection is large. However, it might be practical in the future when the computer technology improves.

7. REFERENCES:

[1] Cross Language Information Retrieval: In Indian Language Perspective by Pretibial Bajpai1 , Rd.ParulVerma.

[2] Cross Language Information Retrieval by Ananthakrishnan R, Natural Language Processing/Language Technology for the Web

[3] 4. Pirkola, A., Hedlund, T., Keskustalo, H., Jarvelin, K.: Dictionary-Based Information Retrieval: Problems, Methods and Research Findings. Information Retrieval 4, 209–230 (2001).

[4] 4.Chinnakotla Kumar Manoj, Ran dive Sagar, Bhattacharyya Pushpak and Damani P. Om “Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007”, in the working notes of CLEF 2007. [5].

[5] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th International Conference on Research and Development in Information Retrieval, pages 49–57, 1996. .