# A SURVEY ON CROSS LANGUAGE INFORMATION RETRIEVAL

Prof. Sanjay K.Dwivedi[1] and Ganesh Chandra[2]
[1-2]Department of Computer Science, BBAU (A Central University), Lucknow

## ABSTRACT

*Now a days, number of Web Users accessing information over Internet is increasing day by day. A huge amount of information on Internet is available in different language that can be access by anybody at any time. Information Retrieval (IR) deals with finding useful information from a large collection of unstructured, structured and semi-structured data. Information Retrieval can be classified into different classes such as monolingual information retrieval, cross language information retrieval and multilingual information retrieval (MLIR) etc. In the current scenario, the diversity of information and language barriers are the serious issues for communication and cultural exchange across the world. To solve such barriers, cross language information retrieval (CLIR) system, are nowadays in strong demand. CLIR refers to the information retrieval activities in which the query or documents may appear in different languages. This paper takes an overview of the new application areas of CLIR and reviews the approaches used in the process of CLIR research for query and document translation. Further, based on available literature, a number of challenges and issues in CLIR have been identified and discussed.*

## KEYWORDS

*Cross language information retrieval, query translation, document translation, corpora based translation, dictionary based translation, machine translation*

## 1. INTRODUCTION

Information retrieval (IR) generally refers to the process where the users search for required information from a large number of documents. Traditional IR systems are implemented mainly for monolingual documents. However, with rapid development of the Internet, the demand for searching information from multi-lingual documents is increasing, which results in the great challenge of how to match the users query written in one language with the documents written in other languages. Therefore, suitable techniques are required to enhance the performance of IR, CLIR & MLIR. The Cross Language Information Retrieval (CLIR) [1] provides a convenient way that can solve the problems of language boundaries, where users can submit queries written in their own language and retrieve documents in another language [2]. For example, a query in Hindi will return the related information description in Chinese language, as given in figure1.

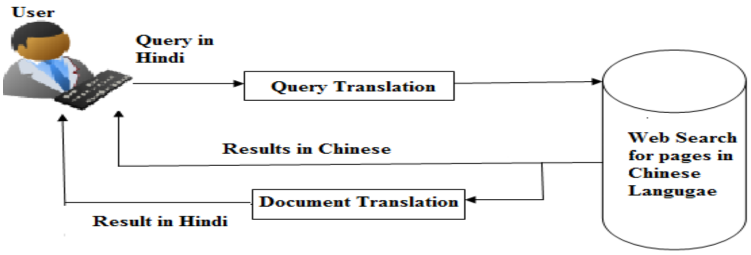

Figure.1 Cross Language Information Retrieval System

DOI: 10.5121/ijci.2016.5113

With the rapid development of Internet technology, globalization of information structure caused the urgent demand to CLIR, because CLIR allows the usage of information exchanges between different languages, remove linguistic disparity between the queries that are submitted and documents that are retrieved using resource over the network, and also decreases the communication cost [3]. The research on information retrieval, came into existence since early 1970s where as experiments for retrieving information across languages were first initiated by Salton in 1973[67]. However, most of the modern research on CLIR started in 90's, and now a day's it has become one of the most important research topic in the area of information retrieval.
An ever active research field, a vast number of researches and studies have been published on CLIR and various issues are addressed in several evaluation forums such as TREC [64], NTCIR [65], and CLEF [66] while each of them cover different languages: TREC includes Spanish, Chinese, German, French, Arabic and Italian; NTCIR includes Japanese, Chinese and Korean and CLEF includes French, German, Italian, Spanish, Dutch, Finnish, Swedish and Russian [4]. The most effective way to solve the problem of language barriers may be achieved through CLIR by using query translation approach, document translation approach, or by using both query and document translation approach.

This paper is organised as follows: section 2, 3, 4 consists of query translation, document translation and dual translation respectively. Section 5, 6, 7 describes the comparative study of the three approaches, challenges of CLIR and CLIR tools respectively. Section 8 describes the application areas of CLIR, section 9 describes CLIR researches in Foreign & Indian languages and section 10 describes the conclusion and future of CLIR.

## 2. QUERY TRANSLATION APPROACH

A major challenge in CLIR is to bridge the language gap between query and documents. Query translation is now serving as a major cross-lingual mechanism in current CLIR systems [5], [40], [70] as shown in figure2. CLIR search engines enable users to retrieve content in a language different from language used to formulate the query. Translation of query has the advantage that the computational effort i.e. time and space, is less as compared with other methods. Query translation has following disadvantages:

(i) Usually a query does not provide enough contexts to automatically find the intended meaning of each term in the query.
(ii) Translation errors affect retrieval performance sensibly.
(iii) In case of searching a multilingual database, query has to be translated into each one of
the languages of database.
In CLIR query translation play an important role that can be achieved by following approaches: dictionary based translation approach, corpora based translation approach and machine translation based approach.

### 2.1. Dictionary Based Translation Approach

In dictionary-based query translation, the query will be processed linguistically and only keywords are translating using Machine Readable Dictionaries (MRD), given in figure3. MRDs are electronic versions of printed dictionaries, either in general domain or specific domain. The use of existing linguistics resources, especially the MRDs, is a natural approach to cross-lingual IR. Translating the query using the dictionaries is much faster and simpler than translating the documents [6], [7], [41]. Some common problems associated with dictionary-based translation [8] are:

(i) Untranslatable words (like new compound words, proper names, spelling variants, and special terms): Not every form of words used in query is always found in dictionary. Some time problem occurs in translating different compound words (formed by combination of new words) due to the unavailability of their proper translation in dictionary [42].

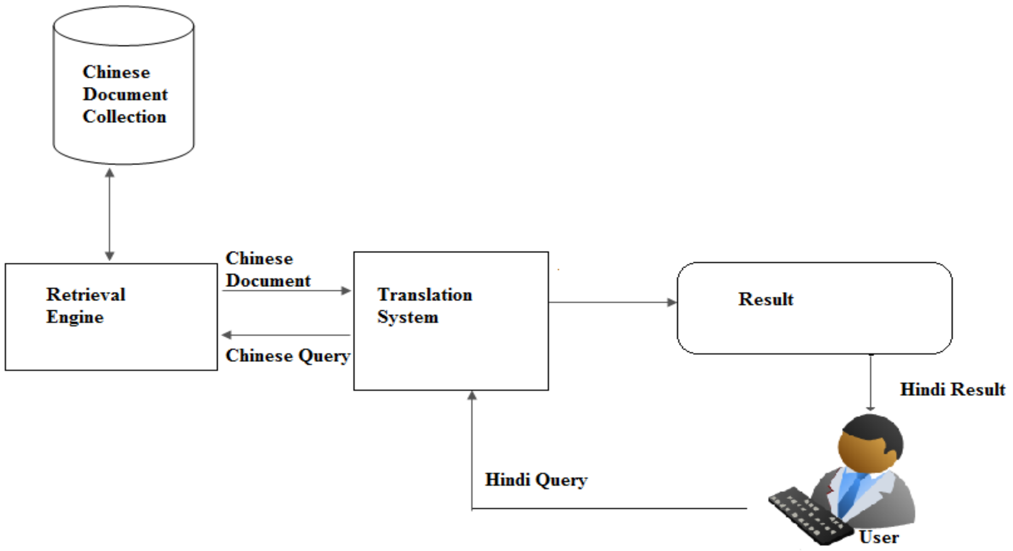

Figure 2: Query Translation

(ii) Processing of inflected words: Inflected word forms are usually not found in dictionaries [43].
(iii) Lexical ambiguity in source and target languages: Relevant forms of lexical meaning for information retrieval are: 1) homonymous and 2) polysemous words. Two words are homonymous; if they have at least two different meanings and senses of words are unrelated e.g. bank (river bank) and bank (financial institution). Polysemous words should have related senses e.g. star in the sky and star. Due to ambiguity in the search keys, matching for retrieving relevant documents may not be successful [46].

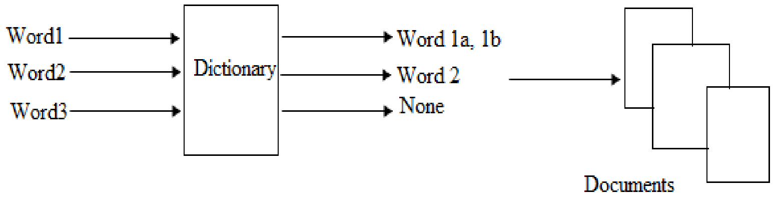

Figure3.Dictionary Based Translation

## 2.2. Corpora Based Translation Approach

Query translation using corpora requires single corpus or many corpuses. Corpora, (plural of corpus) are the systematic collection of naturally occurring language material, such as texts, paragraphs and sentences from one or many languages. In corpus-based methods [9], [12] queries are translated on the basis of multilingual terms extracted from parallel or comparable document collections. A parallel corpus has been used since the early 1990's for translation of given word. A parallel corpus is a collection of texts, each of which is translated into one or more languages other than the original language. Parallel corpora are also used to decide the relationships, such as co-occurrences, between terms of different languages. A parallel corpus is an important kind of source of linguistic meta-knowledge, which forms the basis of techniques such as tokenization, morphological and syntactic analysis [15], [10].

A comparable corpus is one of the important concepts in corpus-based translation study, introduced by Baker [38]. Comparable corpora contain text in more than one language. The texts in each language are not translations of each other, but cover the same topic area, and hence contain an equivalent vocabulary. A good example of corpora is the multilingual news feeds produced by news agencies such as Reuters, CNN, BBC, Xinhua News and BERNAMA. Such texts are widely available on the Web for many language pairs and domains. They often contain many sentence pair that are fairly good translations of each other [11], [12] [13].

## 2.3. Machine Translation Based Approach

Cross-lingual IR with query translation using machine translation [6] seems to be an obvious choice compared to the other two above, as shown in figure4. The advantages of using the machine translation is that it saves time while translating large texts. Manning and Schutze [16] distinguished four different approaches to deal with machine translation: (a) Word-for-word approach, (b) Syntactic transfer approach, (c) Semantic transfer approach, and (d) Interlingual approach. The ultimate goal of CLIR machine translation (MT) systems is to translate queries from one language to another by using a context. Many factors contribute the difficulties of machine translation, including words with multiple meanings, sentences with multiple grammatical structures, uncertainty about what a pronoun refers to, and other problems of grammar.

Many researchers criticize MT-based CLIR approach. The reasons behind their criticisms mostly stem from the fact that the current translation quality of MT is poor. Another reason is that MT systems are expensive to develop and their application degrades the retrieval efficiency (run time performance) due to the lengthy processing times associated with linguistic analysis.

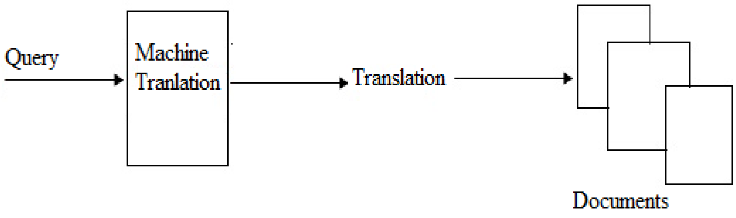

Figure4. Machine Translation

MT based approaches seems to be the ideal solution for CLIR. It is mainly because MT systems translate the sentence as a whole, and the translation ambiguity problem is solved during the analysis of the source sentence. Table 1, describes the differences between various techniques of query translation.

Table1: Comparison between Techniques of Query Translation

| **Parameters** | **Dictionary Based Translation Approach** | **Corpora Based Translation Approach** | **Machine Translation Based Approach** |
|---|---|---|---|
| Ambiguity | High | Low | Low |
| Offline Translation | Possible | Possible | Not possible |
| Working Architecture | Visible as like white box testing. | Visible as like white box testing. | Works similar to black box testing |
| Development | Less expansive | More expensive than | More Expensive |

| | | | |
|---|---|---|---|
| expenses | | DBT | |
| Translation Availability | Highly available in many languages | Available only in few languages | Available only in few languages |

## 3. DOCUMENT TRANSLATION APPROACH

Document Translation [44], [45] can be the most desirable scenario in CLIR, if the purpose is to allow the users to search the documents different from their own language and receive results back in user's language, as given in figure5. In this sense, it is truly a better option which does not require a passive knowledge of the foreign language from the user. In document translation approach, all target languages are translated to the source language [63]. The function of this translation is twofold. First, post translation or 'as-and-when-needed' or 'on-the–fly translation', where documents of any other language being searched by user are translated into user language at query time. IR process mostly uses indexing technique to speed up the searching process of documents. But indexing is not possible in post translation, so this approach is infeasible because it requires more time for translation.

Second, pre translation or 'all together before any query is processed' used to browse through a translated version of an original translation in user language or in a language which user can understand (figure 5). This translation can be called as offline translation. In this approach, documents that are written in different languages are translated to all desired source languages and these documents are indexed before query time. This translation is impossible as a solution for large collection of distributed documents, which are managed by different groups of people, for example internet. Document translation has its own advantages and disadvantages compared to query translation. Some researchers have used it to translate large sets of documents (e.g., Braschler & Schauble, 2001 [17]; Franz, Scott McCarley, & Todd Ward, 2000 [18]; Oard & Hackett, 1998 [19]) since more varied context within each document is available for translation, which can improve translation quality. The document translation approach has certain benefit over query translation. These include the following:

(i) A long document provides more contexts to perform translation, so that terms in the target language can be chosen more accurately.
(ii) Translations errors should not harm retrieval too much, as they are weighted against a whole document.
(iii) The translation effort is done at indexing time, thus getting faster retrieval at run time.
However, there are certain issues with document translation as well, such as:
(i) Much more computational effort is needed to index collections.
(ii) Bad scaling performed in case of more than two languages.

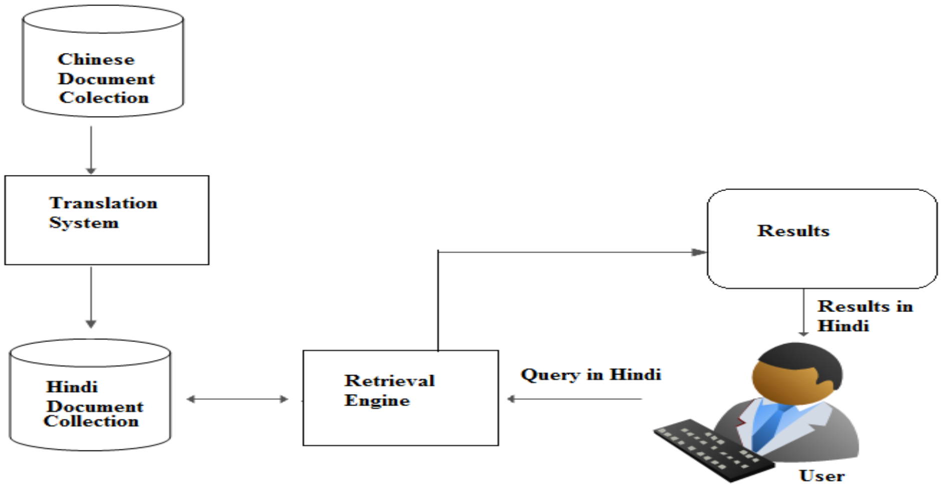

Figure5. Document translation

## 4. Dual translation (both query and document translation approach)

In this approach – both queries and documents are translated into a common representation (figure 6). This approach requires additional storage space for translated documents but provides scalability when same collection of documents is required in multiple languages. One of the examples of such approach is controlled vocabulary systems [20]. These systems represent all documents using a pre-defined list of language-independent concepts, and enforce queries in the same concept space. This concept space defines the granularity or precision of possible searching. The major issue of controlled vocabulary systems is that, non-expert users usually require some training and also require interfaces to the vocabularies in orderable to generate effective queries.
Dual translation approach also called as hybrid translation approach can be performed by pivot language. Direct translation between two languages may not always be possible due to the limitation of translation resources. To perform such type of translation, a resources or a third language is required between these languages, called pivot language. In this process, two types of approaches are possible: either the query or the document is translated first to pivot language, then to the target language; translate both document and query into pivot language as shown in figure 6.

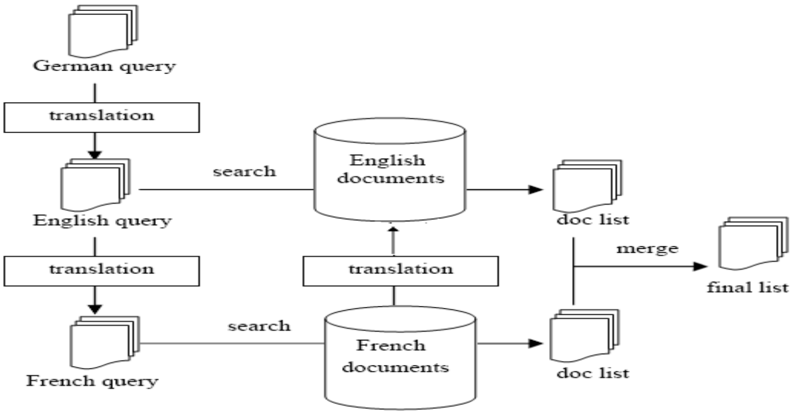

Figure 6. Dual Translation (Pivot Language)

## 5. COMPARATIVE STUDY OF THE THREE APPROACHES

The need for translation has itself been questioned because non-translation based methods of CLIR, such as cognate-matching [21] and cross-language Latent Semantic Indexing [22] have been developed. Document translation into query language or query translation into documents language are the two approaches that couples machines translation and information retrieval. Query translation and document translation approaches are neither equivalent nor mutually exclusive. They are not equivalent because machine translation is not an invertible operation. Query translation and document translation become equivalent only if each word in one language is translated into a unique word in other languages.

Various researches suggest that document translation should be competitive or superior to query translation. Typical queries are short and may contain key words or phrases only. When these are translated inappropriately, the IR engine has no chance to recover. Translating a long document, MT engine offers the many more opportunities to translate key words and phrases. If some of these are translated inappropriately, the IR engine has at least a chance of matching these to query terms. Query translation approach is flexible and allows for more interactions with the user. However, query translation often suffers from the problem of translation ambiguity, and this problem is amplified due to the limited amount of context in short queries. From this perspective, document translation seems to be more capable of producing more precise translation due to richer contexts.

One of the critical aspects of document translation approach is that one has to determine in advance to which language each document should be translated and that all the translated versions of the document should be stored. In a multilingual IR environment, one would desire to translate each document to all other languages. This is impracticable because of the multiplication of document versions and the increase in storage requirement. Once a document is pre-translated into the same language as the query, user can directly read and understand the translated version. Otherwise, a post-retrieval translation is often needed to make the retrieved documents readable by the user (if he/she does not understand the document language).

Query translation and document translation become equivalent only if each word in one language is translated into a unique word in the other languages. Document translation can be performed off-line and on-line but query translation is performed only on-line. Hybrid system that uses both query and document translation are possible because of a trade off between computer resources and quality of translation. Hybrid or dual translation approach provides the relationship between multilingual and the key advantages of these systems are that queries can be expressed and matched unambiguously. In this approach the additional storage space requirement is independent to the number of languages supported. The major problems occurs in this approach are to define the concept space, intermediate representation and conversion of documents into intermediate representation. Differences between two approaches (query translation and documents translation) of CLIR are described in table2. Table3 describes the comparative study of three approaches of CLIR.

Table2: Difference between Query and Document Translation

| Parameter | Query Translation | Document Translation |
|---|---|---|
| Size | Small | Large |
| Language | Prior knowledge of translation language is not required | Prior knowledge of translation language is required |
| Overhead | Low | High |
| Recovery | When these are translated inappropriately, the IR engine has no chance to recover | Chance to recover |
| Ambiguity | Maximum chances of occurring ambiguity | Minimum chances of occurring |

| | | |
|---|---|---|
| | | ambiguity |
| Cost | Low cost | High cost |

Table 3: Comparison of three Translation Approaches

| Parameter | Query Translation | Document Translation | Both Query & Document Translation |
|---|---|---|---|
| Ambiguity | Maximum | Minimum | More than both |
| Additional Storage Space | Not required | Required | Not required |
| Translation time | Less | More than query | More than both |
| Information retrieval | Bilingual | Bilingual | Bilingual and Multilingual |
| Flexibility | Highly | Less | Less |
| Working nature | Can provide interface between two language at a time | Can provide interface between two language at a time | Can provide interface between more than two language at a time |

## 6. CHALLENGES IN CLIR

Queries from users are often too short, which produce more ambiguity in query translation, and reduce the accuracy of the cross language retrieval results. Since the problem of language mismatch in CLIR are more serious than in monolingual IR, it is necessary to exploit techniques for improving the multilingual retrieval performance. In CLIR systems, users often present their query in their native language, and then the system automatically searches documents written in other languages. Therefore, it is a challenge for CLIR to conquer the barrier between the source language (SL) in query sentences and the target language (TL) in documents to be searched. As discussed in the previous section, most CLIR systems utilize MT technology to resolve this problem. As MT research itself has a number of issues (such as accuracy), the research in CLIR also faces critical issues and challenges that must be addressed.

### 6.1. Ambiguity

Ambiguity [15] occurs when words have multiple meaning which also referred to as homonymy or polysemy. Ambiguity in IR are semantic and syntactic in nature, where as ambiguity in CLIR are semantic and lexical. So the probability of occurrence of ambiguity in CLIR is higher than normal IR, due to the availability of different languages [39].

### 6.2. Effective User Feedback

Effective user functionality can be incorporated by the user feedback, about their requirements and information needs. It should also provide readable translations of the retrieved documents to support document selection. System should also provide better support for query formulation and reformulation based on some set of intermediate results.

### 6.3. Complexity in New Applications

Question/Answering is relatively a new stream of Information Retrieval. In Question/Answering end-users throw a question in a form of query and retrieve answers for that. However, challenge is to retrieve answers of English questions in different language.

### 6.4. Specialized Terminology and Proper Nouns

Specialized terminology, such as scientific names, is often difficult to translate and is often found in specialized dictionaries or term banks. Specialized terminology tends to be less ambiguous than regular vocabulary although regular vocabulary can have a specialized meaning when used in a certain subject area.

## 7. CLIR RESEARCHES IN FOREIGN & INDIAN LANGUAGES

India is a multilingual country, with 22 constitutionally recognized languages and 12 scripts. A huge amount of information on Web is available in English. In view of the fact that a small number of people know English in India, others are familiar with Hindi and other local languages. In order to solve the language gap between people of India, many government [14] and private organisations are working on CLIR research field. Some of the prominent organisations working in CLIR [47] field are:

(i) Indian Institute of Technology, Bombay, Maharashtra, India.
(ii) Indian Institute of Technology, Kharagpur, West Bengal, India.
(iii) Indian Institute of Technology, Hyderabad, Telangana, India.
(iv) Anna University- K B Chandrasekhar (AUC-KBC), research centre, Chennai, India.
(v) Indian Statistical Institute, Kolkata, West Bengal, India.
(vi) Jadavpur University, Kolkata, West Bengal, India.
(vii) Centre for Development of Advanced Computing (C-DAC), Pune, India.
(viii) Centre for Development of Advanced Computing (C-DAC), Noida, UP, India.
(ix) Utkal University Bhubaneswar and STDC, Orissa, India.
(x) Department of Information Technology (DIT), New Delhi.
(xi) Microsoft research centre, Bangalore, Karnataka, India.
(xii) Technology Development for Indian Langauge (TDIL), government of India.

The objective of these organisations is to help any user to give a query in one Indian language, can access documents in any other language. Some Foreign organisation that working on CLIR are as follows:

(i) Centre for Intelligent Information Retrieval (CIIR), School of Computer Science, University of Massachusetts Amherst.
(ii) National Institute of Standard and Technology (NIST), U.S. department of research.
(iii) Johns Hopkins University, Baltimore, Maryland.

Some of the prominent researches in Foreign & Indian languages are described in table 4 and table5 respectively as follows:

Table 4: Prominent CLIR research in foreign languages

| Languages | Authors | Year |
|---|---|---|
| English and French | David A. Hull, Gregory Grefenstette [48] | 1996 |
| Japanese and English | Fujii, A., Ishikawa, T.,[49] | 2001 |
| English and Chinese | Su Liu [50] | 2001 |
| Japanese, English and Chinese | Wen- Cheng Lin Hsin- Hsi Chen [51] | 2003 |
| Greek, Latin and Old Norse | Jeffrey A, Rydberg-Cox, Lara Vetter, Stefan M, Ruger Daniel Heesch [52] | 2004 |
| Chinese, Japanese and Korea | Chen-Yu Su, Tien-Chien Lin , Shih- Hung Wu [53] | 2007 |
| English and Chinese | Hsin-Chang Yang, Chung-Hong Lee [54] | 2008 |

Table 5: A brief description of some CLIR research in Indian languages

| Languages | Name of Researcher's | Year |
|---|---|---|
| Hindi, Telugu, English | Prasad Pingali,Vasudeva Verma [56] | 2006 |
| Hindi,Tamil, Telugu, Bengali, Malyalam and English | Jagdeesh, J. , Kumaran, K. [57] | 2007 |
| Telugu, English | Ranbeer Makin, Nikita Pandey, Prasad Pingali, Vasudeva Verma [58] | 2007 |
| English, Hindi | Anurag Seetha, Sujoy Das, M. Kumar [59] | 2007 |
| Hindi, English | Sethuramalingam S, Vasudeva Verma [60] | 2008 |
| English, Hindi, Telugu, Tamil | Manoj Kumar Chinnakotla, Om P. Damani [61] | 2009 |
| English, Telugu, Tamil | P. Sujatha, P. Dhavachelvan, V. Narasimhulu [62] | 2010 |
| Tamil to English | S.Saraswathi & A. Siddhiqaa [55] | 2010 |

## 8. CLIR TOOLS

Over the past few years, research in CLIR has progressed and a many systems have been developed. Some of the prominent systems of CLIR are as follows:

### 8.1 SAPHIRE

The architecture of SAPHIRE system is based on multilingual aspects of UMLS (Unified Medical Language System). In this system a dictionary based approach of CLIR is used [27]. It provides an intelligent healthcare monitoring architecture for high quality health care services with reasonable cost.

### 8.2. KEIZAI

KEIZAI [24] was developed at New Mexico State University and its aim is to provide the Web-based cross language text retrieval system, which searches the documents of Korean or Japanese language on the web for English query. Keizai examines the effectiveness of representing the retrieved documents together with small images, which are called Document Thumbnail Visualizations. The advantage of visualization is to improve the recall and efficiency.

### 8.3. MULINEX

MULINEX system [26] was developed at German Research Center for Artificial Intelligence (DFKI), whose aim is to allow the user to search the collection of multilingual document, supported by an effective combination of linguistic and IR technologies. There are three document categorisation algorithms are used in Mulinex for different tasks: n-gram categoriser for noisy input, k-nearest-neighbour (KNN) algorithm for normal documents, and pattern-based categoriser for every short documents.

### 8.4. MIRACLE

MIRACLE (Maryland Interactive Retrieval Advanced Cross-Language Engine), deals with a combination of statistical and linguistic resources, for monolingual, cross-lingual and multilingual search. In MIRACLE two types of query translation are performed: fully automatic query translation and user assistant query translation [25].

### 8.5. UCLIR

The core process of UCLIR (Unicode Cross-language Information Retrieval system) includes machine translation and standard monolingual information retrieval, which accepts the query in one language and retrieves relevant documents in other language. The UCLIR retrieval system is based on URSA (Unicode Retrieval System Architecture), which is a high-performance text retrieval system that can index and retrieve Unicode texts [28].

### 8.6. KANSHIN

KANSHIN collects and analyzes the multilingual articles of Japanese, Chinese, Korean and English languages [23]. The system provides a various view points for user such as temporal, focal, geographical, and network. It also provides a cross-lingual keyword navigation tool between splog survey tools (called SplogExplorer) and inters language links of Wikipedia.

## 9. APPLICATION AREAS OF CLIR

The core field of information retrieval where research on CLIR is needed for effective results are:

### 9.1. Medical

A number of resources available on Web provide the public and healthcare professionals with the most up-to-date findings in medical research, such as PubMed [68] and MedlinePlus [69].

Table6: List of some prominent CLIR tools from 1990's

| Name | Language | Developed At | Year |
|---|---|---|---|
| SAPHIRE [27] | English, German, French, Russian, Spanish, and Portuguese. | School of Medicine Oregon Health Sciences University Portland, OR, USA by William R. Hersh, M.D., Laurence C. Donohoe, M.L.I.S | 1998 |
| KEIZAI [24] | English , Japanese & Korean | Computer Research Lab New Mexico State University, Las Cruces USA by William Ogden & James Cowie et.al | 1999 |
| MULINEX[26] | French, German and English | German research Centre for artificial Intelligence by Capstick et.al | 2000 |
| MIRACLE [25] | English , French, German, Spanish, Cebuano and Hindi | Spanish University (Universal Politecnica de Madrid, Universidad Autonoma de Madrid & Universidad CarlosIII de Madrid) by Julio, Sara et.al | 2003 |
| UCLIR [28] | English , Japanese Korean, Arabic | Computing Research Laboratory at New Mexico State University by Ahmed Abdelali et.al | 2004 |
| KANSHIN [23] | Japanese, Chinese , Korean & English | University of Tokyo Japan by Tomohiro Fukuhara et.al. | 2005 |

Medline Plus is a Web-based consumer health information resource, made available by the National Library of Medicine. PubMed first released in 1996, is a free search engine for accessing the Medline database of life sciences and biomedical topics.

Most of the high level quality resources that are freely available and unlimited for users all around the world are available only in English language. Therefore Non-English users encounter a great language barrier when trying to access medical information from these websites such are also not familiar with medical terminology even in their first language (native language). So there is a big platform for researcher to work on medical information retrieval system, in order solve the problem of language barrier [29].

### 9.2. Multimedia

Multimedia Information Retrieval (MMIR or MIR) is a hot research discipline whose objective is to extract the semantic information from multimedia data sources such as audio, video, and image [1]. MMIR implies that multiple channels are employed for the understanding of media content, each of these channels are described by media-specific features transformations [30].

The first version of the Multilingual Multimedia Information Retrieval (MMIR) prototype involves short videos in the domain of news, that are selected from online web TV channels, from UGC portals, or from online news agencies [31]. There has so far been very little work in the area of Cross-Language Multimedia Information Retrieval (CLMIR). This is an important future research topic as the growth of multilingual and multimedia document collections is likely to lead inevitably to the growth of multilingual multimedia collections [32], [33].

### 9.3. Mobile Network

This research proposes a Cross-Lingual Information Retrieval approach that is used to search Internet resources for appropriate content and summarize it into another form using the content specification meta-language. This content is then mapped to the target language [34].

### 9.4. Video Question/Answering system

Question/Answering on multi-media is a new research issue in recent years. The cross-language QA system have some fundamental problems like video processing, i.e. video Optical Character Recognition (OCR) [32], [35], [36] and video segmentation.

### 9.5. Enterprise Competition

Along with the economic globalization, the information resource in a modern society becomes an important element for modern enterprises competition. CLIR is introduced to the enterprise competitive intelligence collections can effectively resolve the low recall and veracity rate of intelligence collections to some extent and promote the development of CLIR in the enterprise competition intelligence [37].

## 10. CONCLUSION AND FUTURE OF CLIR

Cross-lingual IR provides new mirror in searching documents through multitude varieties of languages across the world and it can be the baseline for searching not only between two languages but also in multiple languages. Today, most of the cross-lingual researches involved only few famous languages like English, Hindi, Spanish, China and French. Research on languages has increases the development of country. As the world becomes more connected by technology, cross language IR in every language is needed. CLIR is a multidisciplinary area that has been increasingly gaining more attention from the research community. Despite recent advances and new developments, there are still many aspects to be explored.

In Indian context, which is one of the hotspots of linguistic diversity (350 languages) in the globe, and the fact that a dominant language of one region may be a language of a linguistic minority in other region, cross languages information retrieval systems would play a very important role in allowing the people to go through the documents and literatures of other languages thus breaking the language barrier. We work out here to give a broad overview of the speedy demanding work in the field of CLIR by exploring its aspiration, difficulties, basic tools, major works and future research goals. In reviewing this information, it becomes possible to gain a larger picture of the CLIR field.

## REFERENCES

[1] Gaillard, B., Bouraoui, J. L., De Neef, E. G., & Boualem, M., (2010, May), "Query expansion for cross language information retrieval improvement", In Research Challenges in Information Science (RCIS), 2010 Fourth International Conference on (pp. 337-342). IEEE.

[2] Pigur, V. A., (1979, "Multilanguage information-retrieval systems: Integration levels and language support", Automatic Documentation and Mathematical Linguistics, 13(1), 36-46.

[3] Peng, Qu, Lu, Li & Zhang lili, (2008) "A Review of Advanced Topics in Information Retrieval", Library and Information Service, Vol.52.No 3. pp. 19-23.

[4] Ahmed, F., & Nurnberger, A., (2012), "Literature review of interactive cross language information retrieval tools", Int. Arab J. Inf. Technol., 9(5), 479-486.

[5] Narasimha Raju, B. N. V., Bhadri Raju, M. S. V. S., & Satyanarayana, K. V. V. (2014, December). Translation approaches in Cross Language Information Retrieval. In Computer and Communications Technologies (ICCCT), 2014 International Conference on IEEE, pp. 1-4.

[6] Carley Mc & S.J. (1999, June), "Should we translate the documents or the queries in cross-language information retrieval?" In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, pp. 208-214.

[7] Aljlayl, M., & Frieder, O., (2001, October), "Effective arabic-english cross-language information retrieval via machine-readable dictionaries and machine translation", In Proceedings of the tenth international conference on Information and knowledge management ACM, pp. 295-302.

[8] Pirkola, A., Hedlund, T., Keskustalo, H., & Jarvelin, K., (2001), "Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings," Information Retrieval, 4(3/4), pp. 209-230.

[9] Picchi, E. & Peters, C, (2000) "Cross-language information retrieval: a system for comparable corpus querying, in Cross-Language Information Retrieval," G. Grefenstette, Editor. 2000, Kluwer Academic Publishing: Massachusetts. pp. 81-90.

[10] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 496). Cambridge: Cambridge University Press.

[11] Munteanu, D.S. and D. Marcu. (2005) "Extracting parallel subsentential fragments from non-parallel corpora", In Proceedings of the 21st international Conference on Computational Linguistics. Sydney, Australia: Association of Computational Linguistics.

[12] Landauer, T.K. & Littman, M.L.,(1990) "Full Automatic Cross Language Document Retrieval using Latent Semantic Indexing" , In Proc. of the 6th Conference of UW Center for New OED and Text Research, pp. 31–38.

[13] Sheridan, P., & Ballerini, J.P., (1996) "Experiments in Multilingual Information Retrieval using the SPIDER System", In Proc. of the 19th Annual International ACM SIGIR, pp. 58–65.

[14]Sourabh, K., (2013) "An Extensive Literature Review on CLIR and MT activities in India", International Journal of Scientific & Engineering Research.

[15] Chandra, G.; Dwivedi, S.K., "A Literature Survey on Various Approaches of Word Sense Disambiguation," in Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium on , vol., no., pp.106-109, 7-8 Dec. 2014.

[16] Manning, C.D., & Schutze, H., (1999) Foundations of statistical natural language processing. MIT Press.

[17] Braschler, M., & Schauble, P., (2000), "Using corpus-based approaches in a system for multilingual information retrieval." Information Retrieval, 3, pp. 273–284.

[18] Franz, M., Scott McCarley, J., & Todd Ward, R. (2000), Ad hoc, cross-language and spoken document information retrieval at IBM. In Proceedings of TREC-8. Gaithersburg, MD: National Institute of Standards and Technology. Available: http://trec.nist.gov/pubs/.

[19] Oard, D. W., & Diekema, A. R., (1998) "Cross-language information retrieval", Annual Review of Information Science and Technology, 33, pp. 223–256.

[20] Oard D and Dorr B. (1996), "A Survey of Multilingual Text Retrieval", Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

[21] Buckley, M. Mitra, Wals, J. & Cardie, C., (1998), "Using clustering and superconcepts within SMART: TREC-6", In E.M. Voorhees and D.K. Harman, editors, The 6th Text REtrieval Conference (TREC-6).

[22] Dumais, S. T., Letsche, T. A., Littman, M. L., & Landauer, T. K., (1997, March) "Automatic cross-language retrieval using latent semantic indexing", In AAAI spring symposium on cross-language text and speech retrieval, Vol. 15,pp. 21.

[23] Fukuhara, T., Kimura, A., Arai, Y., Yoshinaka, T., Masuda, H., Utsuro, T., & Nakagawa, H., (2008, April) "KANSHIN: A cross-lingual concern analysis system using multilingual blog articles", In Information-Explosion and Next Generation Search, 2008. INGS'08. International Workshop on IEEE, pp. 83-90.

[24] Ogden, W., Cowie, J., Davis, M., Ludovik, E., Nirenburg, S., Molina-Salgado, H., & Sharples, N. (1999, September). Keizai: An interactive cross-language text retrieval system. In Proceeding of the MT SUMMIT VII workshop on machine translation for cross language information retrieval, Vol. 416.

[25] He, D., Oard, D. W., Wang, J., Luo, J., Demner-Fushman, D., Darwish, K., & Leuski, A., (2003), "Making miracles: Interactive translingual search for cebuano and hindi", ACM Transactions on Asian Language Information Processing (TALIP), 2(3), 219-244.

[26] De Luca, Ernesto William, Ernesto William, Stefan Hauke De Luca, Andreas Nürnberger, and Stefan Schlechtweg., (2006) "MultiLexExplorer-Combining Multilingual web Search with Multilingual Lexical Resources", In Proceedings of the combined workshop on Language-Enabled educational technology and Development and evaluation of Robust Spoken Dialogue Systems, Germany, pp. 17-21.

[27] Hersh, William R., and Larry C. Donohoe, (1998) "SAPHIRE International: a tool for cross-language information retrieval.", Proceedings of the AMIA Symposium. American Medical Informatics Association, pp673.

[28] Abdelali, A., Cowie, J. R., Farwell, D., & Ogden, W. C., (2004) "UCLIR: a Multilingual Information Retrieval Tool", Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial, 8(22), pp103-110.

[29] Lu, Wen-Hsiang, et al. (2008), "Using Web resources to construct multilingual medical thesaurus for cross-language medical information retrieval." Decision Support Systems 45.3: 585-595.

[30] Arrieta, K., Leturia, I., Iturraspe, U., de Ilarraza, A. D., Sarasola, K., Hernáez, I., & Navas, E. (2008, December). AnHitz, development and integration of language, speech and visual technologies for Basque. In Universal Communication, 2008. ISUC'08. Second International Symposium on IEEE, pp. 338-343.

[31] Wu, Y. C., Chang, C. H., & Lee, Y. S., (2004, December) "CLVQ: Cross-language video question/answering system", In Multimedia Software Engineering, 2004. Proceedings. IEEE Sixth International Symposium on IEEE, pp. 294-301.

[32] Lin, C. J., Liu, C. C., & Chen, H. H., (2001) "A simple method for Chinese video OCR and its application to question answering", Computational linguistics and Chinese language processing, 6(2), pp11-30.

[33]Wu, V., Manmatha, R., & Riseman, E. M., (1999) "Textfinder: An automatic system to detect and recognize text in images", IEEE Transactions on Pattern Analysis & Machine Intelligence, (11), pp1224-1229.

[34] Kumar, B. A. (2012, January). Profound survey on cross language information retrieval methods (CLIR). In Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on IEEE, pp 64-68.

[35] Shim, J. C., Dorai, C., & Bolle, R., (1998, August) "Automatic text extraction from video for content-based annotation and retrieval", In Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on IEEE, Vol. 1, pp. 618-620.

[36] Yang, H., Chaisorn, L., Zhao, Y., Neo, S. Y., & Chua, T. S. (2003, November). VideoQA: question answering on news video. In Proceedings of the eleventh ACM international conference on Multimedia ACM, pp. 632-641.

[37] Xiang-jun, L., Hong-Na, S., Li-Wen, Y., & Xiao-Kun, Y. (2009, May), "Research of Enterprise Competitive Intelligence Collection System Based on Cross-Language Information Retrieval", In Electronic Commerce and Security, 2009. ISECS'09. Second International Symposium on IEEE, Vol. 1, pp. 601-604.

[38] Fernandes, Lincoln Paulo. (2006), "Corpora in Translation Studies: revisiting Baker's tipology." Fragmentos: Revista de Língua e Literatura Estrangeiras, 30.

[39] Diekema, Anne R., (2003) "Translation events in cross-language information retrieval: lexical ambiguity, lexical holes, vocabulary mismatch, and correct translations".

[40] Wu, D., & He, D., (2010, December) "A study of query translation using google machine translation system", In Computational Intelligence and Software Engineering (CISE), 2010 International Conference on IEEE, pp. 1-4.

[41] Seetha, A., Das, S., & Kumar, M., (2007, December) "Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method", In Information Technology,(ICIT 2007). 10th International Conference on IEEE, pp. 56-61.

[42] Pfeifer, U., Poersch, T., & Fuhr, N. (1996) "Retrieval effectiveness of proper name search methods", Information Processing & Management, 32(6), pp 667-679.

[43] Fluhr, C., Schmit, D., Ortet, P., Elkateb, F., Gurtner, K., & Radwan, K. (1998). Distributed cross-lingual information retrieval. In Cross-Language Information Retrieval, Springer US, pp. 41-50).

[44] Croft, W. B., Turtle, H. R., & Lewis, D. D. (1991, September), "The use of phrases and structured queries in information retrieval", In Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 32-45.

[45] Buckley, C., Singhal, A., Mitra, M., & Salton, G., (1995, November), "New retrieval approaches using SMART: TREC 4", In Proceedings of the Fourth Text Retrieval Conference (TREC-4), pp. 25-48.

[46] Lyons, J, (1981), "Language and Linguistics: An introduction", Cambridge University Press.

[47] Nagarathinam, A., & Saraswathi, S. (2011), "State of Art: Cross Lingual Information Retrieval System for Indian Languages", International Journal of Computer Applications, 35 (13).

[48] Hull, D. A., & Grefenstette, G., (1996, August), "Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval", In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval ACM, pp. 49-57.

[49] Fujii, A., & Ishikawa, T., (2001, March) "Evaluating multi-lingual information retrieval and clustering at ULIS", In Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization. Tokyo. Japan, pp. 5-144.

[50] Su Liu.: ECIRS, (2001), "An English-Chinese Cross-language Information-retrieval System", In: IEEE International Conference on Systems, Man, and Cybernetics, Volume 2.

[51] Lin, W. C., & Chen, H. H., (2003) "Description of NTU approach to NTCIR3 multilingual information retrieval", In Proceedings of the third NTCIR workshop.

[52] Jeffrey A. Rydberg-Cox, Lara Vetter, Stefan M. Ruger, Daniel Heesch, (2004) "Cross-lingual searching and visualization for greek and latin and old norse texts", In: JCDL Vol. 383.

[53] Su, C. Y., Lin, T. C., & Wu, S. H. (2007, May). Using Wikipedia to translate OOV terms on MLIR. In Proceedings of NTCIR-6 Workshop Meeting (pp. 109-115).

[54] Yang, H. C., & Lee, C. H., (2008, November) "Multilingual Information Retrieval Using GHSOM", In Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on IEEE, Vol. 1, pp. 225-228.

[55] Dr. Saraswathi, S., Siddhiqaa, Asma, Kalaimagal, M, K., and Kalaiyarasi M (2010, April) BiLingual Information Retrieval System for English and Tamil, Journal Of Computing, 2, 4, 85-89.

[56] Pingali, P., & Varma, V. (2006, September) "Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006", In Working Notes of Cross Language Evaluation Forum.

[57] Jagadeesh, J. & Kumaran, K, (2007), "Cross-Lingual Information Retrieval System for Indian Languages.In: Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF, pages: 80-87.

[58] Makin, R., Pandey, N., Pingali, P., & Varma, V., (2007) "Approximate string matching techniques for effective CLIR among Indian languages", In Applications of Fuzzy Sets Theory, Springer Berlin Heidelberg, pp. 430-437.

[59] Seetha, A., Das, S., & Kumar, M. (2007, December). Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method. In Information Technology,(ICIT 2007). 10th International Conference on IEEE, pp. 56-61.

[60] Sethuramalingam, S., & Varma, V. (2008). IIIT Hyderabad's CLIR experiments for FIRE-2008. In The working notes of First Workshop of Forum for Information Retrieval Evaluation (FIRE), Kolkata.

[61] Chinnakotla, M. K., & Damani, O. P., (2009, August) "Experiences with english-hindi, english-tamil and english-kannada transliteration tasks at news 2009", In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration ,Association for Computational Linguistics, pp. 44-47.

[62] Sujatha, P., Dhavachelvan, P., & Narasimhulu, V., (2010) "Evaluation of English-Telugu and English-Tamil Cross Language Information Retrieval System using Dictionary Based Query Translation Method", International Journal of Computer Science and Information Security, 8(2), 314-319.

[63] Ramanathan, A. (2003). State of the Art in Cross-Lingual Information Retrieval.VIVEK-BOMBAY-, 15(2), 16-22.

[64] Voorhees, Ellen M., & Donna K. Harman, eds, (2005), "TREC: Experiment and evaluation in information retrieval. Vol. 63. Cambridge: MIT press, Vol. 63.

[65] Peters, Carol, (2001), "Cross-language information retrieval and evaluation", Berlin Heidelberg: Springer.

[66] Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., & Petras, V. (2006) "Geo CLEF: the CLEF 2005 cross-language geographic information retrieval track overview ,Springer Berlin Heidelberg, pp. 908-919.

[67] Salton, G., (1973), "Experiments in multi-lingual information retrieval", Information Processing Letters, 2(1), 6–11.

[68] Chang, D. T. H., Weng, Y. Z., Lin, J. H., Hwang, M. J., & Oyang, Y. J., (2006) "Protemot: prediction of protein binding sites with automatically extracted geometrical templates", Nucleic acids research, 34(suppl 2), W303-W309.

[69] Shatkay, Hagit, (2005) "Hairpins in bookstacks: information retrieval from biomedical text", Briefings in Bioinformatics 6.3 (2005): 222-238.

[70] Oard, D. W., He, D., & Wang, J. (2008). User-assisted query translation for interactive cross-language information retrieval. Information Processing & Management, 44(1), 181-211.

## AUTHORS

**Prof. Sanjay K. Dwivedi** is working as Professor & Head, Department of Computer Science at Babasaheb Bhimrao Ambedkar Central University, Lucknow, India. His research interest is in Artificial Intelligence, Information Retrieval, Web Mining, NLP and WSD. He has published number of research papers in reputed journals and conferences. He is approachable at skd200@yahoo.com.

**Ganesh Chandra** received his M. Tech. degree from KNIT Sultanpur, U.P. India and is pursuing his Ph.D. degree in Department of Computer Science at Babasaheb Bhimrao Ambedkar Central University, Lucknow, India. His major research interests include Information Retrieval, WSD and Machine Translation. He has published papers in various international journals and conferences. He is approachable at ganesh.iiscgate@gmail.com.