# Survey on Variants of Cross Language Information Retrieval

S. Vaishnavi
Assistant Professor, Department of IT
G. Narayanamma Institute of Technology
& Science
Hyderabad, TS,India
*e-mail:sadula_vaishnavi@yahoo.co.in*

Dr. Anitha Chepuru
Associate Professor, Department of IT
G. Narayanamma Institute of Technology
& Science
Hyderabad, TS,India
*e-mail:anithachepuru@gmail.com*

***Abstract*—**To provide the user with relevant information, is the utmost goal of information retrieval.As the information on web is available in many languages, retrieval need not be restricted to single language. We can place the query in any language and also can search for information in documents represented in any language. In this paper different flavors of retrieval with respect to language like monolingual,bilingual,cross lingual and multilingual are summarized. Different types of resources available to perform the search are also described.

***Keywords****-Cross language information retrieval, translation techniques, parallel corpora, comparable corpora, bi-lingual dictionary.*

*****

## I. Introduction

The web is very vast and it contains information regarding any field or object. Anything you ask it will give you. The World Wide Web has revolutionized the access to information, making just about everything available within a few seconds. So now a days any person who has access to smart phone or computer with internet is using these devises to browse for any type of information or queriesregarding any kind of knowledge like current technologies, new releases of products, news, kids project works, cooking, entertainment, movie reviews, shopping news anything.

To provide the user with relevant information is the purpose of information retrieval. For example, when user searches for information,the results should include documents containing this information either in text or audio or multimedia. World Wide Webnot only provides easy access to required information but even allows retrievalin other languages. i.e In countries like India, where many languages are spoken and understood the search for information need not restrict to the local language of the user, but can be extended to other languages corpus to give required results.

## II. Variants of Information Retrieval

The variants of the Information Retrieval are:-

- Mono-Lingual Information Retrieval
- Bi-Lingual Information Retrieval(BLIR)
- Multi-Lingual Information Retrieval(MLIR)
- Cross-Lingual Information Retrieval(CLIR)

If the retrieval is done using single language. i.e If the query and the corpus are represented in same language it is called**Monolingual information retrieval**. It is the general type of retrieval mostly done using English language.English is the dominant language used in information retrieval for representing queries and corpus.As the data in the web is increasing exponentially and each day many documents are uploaded to web and they are represented not only in one but many different languages as parallel corpus.So there is much information even available in other languages other than English.This internationalization in www had made a important role to shift frommono to bi and multilingual information retrieval.

If the query is given in one language and accessing information is done in language other than that of the query then it is called **Bi-Lingual information retrieval**, for example accessing document in Telugu while using English language for query. This type of retrieval requires source language to be translated into the language of query in order to give results. We need good bilingual dictionary or machine translation programfor Bilingual information retrieval. The limitations of this retrieval are most dictionaries do not include the entire vocabulary of any given language, and one needs to decide which, if any of the proposed words is the right one. SoBilingual Information Retrieval requires a Multiple-Level approach to overcome the issues of translation.

If the user is using more than one language to ask query and also access relevant information in multiple languages then the retrieval is called **Multi lingual information retrieval (MLIR)[1].**Multi lingual information retrieval not only has to take care about access from different languagesbut also about different formats, e.g. PDF, HTML, PHP, etc. It should also take care about rapidly changing dynamic pages and also about proper indexing so that valid and most relevant pages may not be missed.

The task of retrieving documents across languages is described as Cross-language Information Retrieval (CLIR)[1]. In CLIR queries are given in one language and hunt of the query for relevant documents is made in one or more different languages.

The term "cross-language information retrieval" has many synonyms, cross-lingual information retrieval, translingual information retrieval, multilingual information retrieval.Both MLIR and CLIR are sub-fields of information retrieval. MLIR is multilingual with respect to both the query and retrieval, but CLIR is multilingual only with respect to access to relevant information.

Though cross language information retrieval is most prevalent but still Monolingual information retrieval play a very major role in the field of information retrieval, as it is basis for cross-language information retrieval techniques.We need to have effective tools and techniques of monolingual retrieval to develop tools and system for bi and multilingual information retrieval. As each language is different and as such, needs a different approach, developing tools for different languages is challenging.what the user assumes as relevant information also plays very important role to retrieve information in languages other than that of the query as each language has its own interpretation.This gives rise to the issues in cross-language information retrieval (CLIR)

There are four major modules in most CLIR engines as follows:

- Pre-translation module,
- Translation module,
- Post-translation module and
- Information retrieval module.

### A. Pretranslation Module

The pre-translation module is responsible for identifying, extracting, and processing suitable linguistic units present in the source text. This activity can be broken down into four separate activities - tokenization, stop word removal, stemming, and term expansion.

### B. Translation Module

The translation module is the main module of the CLIR process .In order to enable the user access relevant information if the document is not in language of their query, the CLIR has to deal with the translation technique(s) for either of the query or the document.
This module will follow one of two general approaches to translate:

- Direct translation
- Indirect translation.

Furthermore, each of the translation module can operate in one of the following three modes.

*1) Query Translation:*The query is translated into the language in which the document collection is written[5].

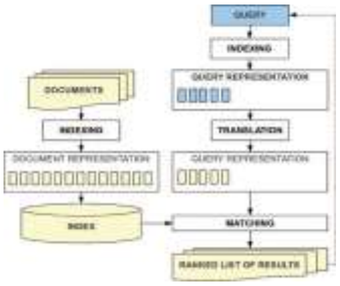

Figure 1. CLIR Using Query Translation

*2) Document Translation:*At indexing time, the documents are translated into the same language as the query[5].

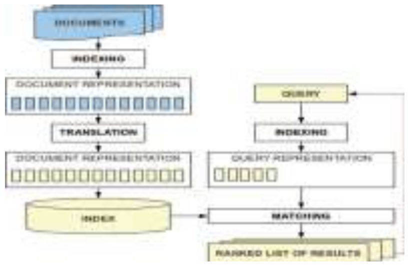

Figure 2. CLIR Using Document Translation

*3) Dual Translation:*Both the query and the document collection are translated into a third language (or semantic space) to enable comparison[5].

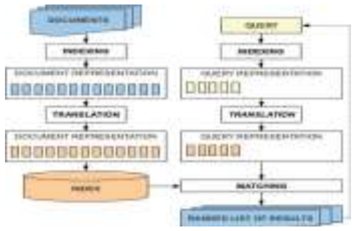

Figure 3. CLIR Using Dual Translation

### C. Post-translation Module

The purpose of the post-translation module is to shape the output of the translation module into the final product. This process may involve expanding the translated text or reweighting individual terms.

### D. Information Retrieval Module

This module does the actual matching of query representations against document representations and ranking the results.

## III. Resources Available for CLIR

The simplest way to search for the information is to scan every item in the database and when the need to translate the languages being used arises, then there will be a need of developing Cross Language Information Retrieval systems .To do so, most CLIR systems use various translation techniques. Based on the translation resources the techniques used for cross language information retrieval are:

- Dictionary-based CLIR techniques
- Parallel corpora based CLIR techniques
- Comparable corpora based CLIR techniques
- Machine translator based CLIR techniques
- Ontology based CLIR techniques
- Transitive based CLIR techniques

### A. *Dictionary-based*

Aconventional Dictionary from query language to document language is used in this technique to map the terms and retrieve information in other language(s) than the one used for the query.

### B. *Parallel corpora*

Parallel corpora ismade bytranslating same text into two or multiple languages. In this technique information is retrievedfrom the so-called parallel corpora which are previously translated.This techniqueeliminates the risk of mistranslations and other problems associated with the dictionary-based technique. But the resultswe get from parallel corpora are limited.

### C. *Comparable corpora*

It works similar to parallel corpora. The difference is here it uses comparable corpora. Comparable corpora contain text in multiple languages which is not the translation of same text but rather deals with the same subject. So the vocabulary can be similar which will help in retrieval.

### D. *Machine translation*

Machine translation is actually very useful and quite reliable too under condition that it is used properly. They use softwares to translate.

### E. *Ontology Based*

In this technique user takes the meaning of words and its context to decide which of the word is correct translation for the query term.

### F. *Transitive and Triangulation Methods*

Transitive and triangulation methods use other languages that have sufficient lexicon resources to accommodate for the lack of bilingual translation resources between a language pair. Transitive methods use a medium language to bridge the translation gap between two languages. Triangulation Methods translations from two different transitive routes are used to extract the translations in a third language.

## IV. ISSUES TO BE TAKEN CARE IN CLIR

CLIR has to consider many issues while translation.It has to take care about, **Translation of phrases**Where each phrase has to be considered entirety, rather than individual word-for-word translation[2].**Disambiguation Techniques**When the query terms can be mapped to many words in translated language with different interpretational meaning, then decision of which of them is the correct mapping for the actual query term is also a crucial issue.As these unwanted or wrongly mapped terms may effect recall and precession rates of relevant results. Disambiguation techniques can be improved using part-of-speech tagging, parallel-corpus based techniques, and query expansion techniques.**Word inflection**can also become a problem in query translation. Example the word can be compute or computing but the words represent similar meaning[2].Lemmatization can be used to solve this problem, where everyword has to be cutshort to its uninflected form or lemma. We can also use stemming, where by removing the word endings from different grammatical forms of a word,a reduced common shorter form called a stem is formed. **Compound words** is a word formed by combination of two or more words where most of the times they occur together. Which can be actually translated as individual words,but if individual terms cannot be translated. Then whole compound word has to be translated. For which mappings may be difficult. **Proper names** and **spelling variants**issues also have to be considered at the time of translation.**Special terms** are also has to be considered as they have specific notation or meaning which may not be available in dictionaries.

## V. CLIR FOR MINORITY LANGUAGES

Though there are many resources available for cross language information retrieval .Effective and readymade resources like parallel corpora are not available for minority languages mostly like local languages of Indian country. For such languages relevant results have to be obtained from dominant language which should be converted into requested language using dictionary based more effectively[10].It is been shown that if N-gram technique is used along with bilingual dictionary for English to Telugu translation the results are good.

TABLE I. LANGUAGE IDENTIFIER ACCURACY

| Language | Language |
|---|---|
| Telugu | 98% |
| Hindi | 99% |
| Tamil | 99% |
| Punjabi | 97% |
| Marati | 98% |
| Bengali | 99% |

Cross-language information retrieval (CLIR) is being focused a lot in the past two decades as the internet usage has become quiet common by any kind of user for any type of request. The consortium of academic, research institutions and industry partners had executed ***Cross Language Information Access Portal for Indian Languages*** project . The languages involved in it are: Bengali, Hindi, Marathi, Punjabi, Tamil, Telugu, Assamese, Oriya and Gujarati.The National Institute for Informatics (NII) of Japan held the first NII Testbeds and Community for Information access Research (NTCIR) which concentrated on Asian languages. Dominant languages gained interest but still focus towards minority languages is necessary.

The language barrier can be resolved using Cross Language Information Retrieval System which makes domain information accessible to all users irrespective of language & region. Cross-language image retrieval is still presenting a major challenge for the IR researchers, however, it also presents a major opportunity.

## REFERENCES

[1] Miss RekhaWarrier, MrsSharvari S. Govilkar “A SURVEY ON VARIOUS CLIR TECHNIQUES”April 2015.

[2] MustafaAbusalah, JohhTait, MichaelOakes”Literature Review of Cross LanguageInformationRetrieval”May 2014

[3] Peishan Tsai “An Introduction to Cross-Language Information Retrieval Approaches”

[4] Monika Sharma1, SudhaMorwal” A Survey on Cross Language InformationRetrieval” Feb 2015

[5] DONG ZHOU,MARK TRURAN,TIM BRAILSFORD, VINCENT WADE, HELEN ASHMAN, “Translation Techniques in Cross-Language Information Retrieval”Nov,2012,ACM ,vol1

[6] KVN Sunitha, N Kalyani”A Novel approach to improve rule based Telugu morphological analyzer”,IEEE,Dec 2009

[7] RaziehRahimi,AzadehShakery” Online Learning to Rank for Cross-Language Information Retrieval”ACM Aug 2017

[8] Dong Zhou,SéamusLawless,JianxunLiu“Query expansion for personalized CLIR retrieval”IEEE Nov,2015

[9] DasuUjjwalPrakharRastogiSirilSiddhartha”Analysis of retrieval models for cross language information retrieval”IEEE,jan.2016

[10] Cross language information access in teluguby Vasudeva Varma, Aditya MogadalaMogadala, V. Srikanth Reddy, Ram Bhupal Reddy in *Siliconandhrconference (Global Internet forum for Telugu)2011*

[11] Cross-Language Information Retrieval by Jian-Yun NieUniversity of Montreal

[12] B.N.V Narasimha Raju ,M S V S Bhadri Raju ,K V VSatyanarayana”Translation approaches in Cross Language Information Retrieval”IEEE,mar 2015

[13] Jian-Yun Nie, Michel Simard, Pierre Isabelle, Richard Durand” Cross-Language Information Retrieval asedonParallel Texts and Automatic Mining of ParallelTexts from the Web”

[14] Ananthakrishnan R “Cross-Language Information Retrieval (CLIR)

[15] N. Kalyani , Dr. K. V. N. Sunitha”Isolated Word Recognition using Morph – Knowledge for Telugu Language”

[16] THosmasTalvensaari”Comaparable corpora in CLIR”

[17] Douglas W_ Oard”A comparative study of query and document translation for CLIR”

[18] N. Kalyani , Dr. K. V. N. Sunitha” “Syllable analysis to build a dictation system in Telugu language”

[19] Fatiha Sadat, Masatoshi Yoshikawa, ShunsukeUemura“Bilingual Terminology Acquisition from Comparable Corpora and Phrasal Translation to Cross-Language Information Retrieval”

[20] Douglas w.Oard”Document translation for cross language information retrieval at university of Maryland”

[21] Atsushi fujii,Tetsuyaishikawa “Cross language information retrieval for technical documents”

[22] JiangpingChen”Cross language search:the case of google language tools”