ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A SURVEY ON VARIOUS CLIR TECHNIQUES

Mrs.Rekha Warrier*, Mrs.Sharvari .S.Govilkar
*[1]Department of Computer Engineering, Mumbai University
PIIT, Navi Mumbai, India
*[2] Department of Computer Engineering, Mumbai University
PIIT, Navi Mumbai, India

Abstract

Information retrieval (IR) system aims to retrieve relevant documents to a user query where the query is a set of keywords. Cross-language information retrieval (CLIR) is a retrieval process in which the user fires queries in one language to retrieve information from another language. The growing requirement on the Internet for users to access information expressed in language other than their own has led to Cross Language Information Retrieval (CLIR) becoming established as a major topic in IR.

Keywords: Cross Language Information Retrieval,Ontology, multilingual, dictionary-based translation, corpora, machine translation

Introduction

The area of Information Access has evolved to perform many sophisticated tasks such as the information retrieval, question answering tasks, summarization, multimedia information retrieval, text mining and clustering and Web information retrieval . Cross-language Information Retrieval (CLIR) can be described at an abstract level as the task of retrieving documents across languages. This deals with firing queries in one language and retrieving documents in one or more different languages. The variants of the IR are

1) **BLIR**(Bi-Lingual Information Retrieval)
2) **CLIR**(Cross-Lingual Information Retrieval) and
3) **MLIR**(Multi-Lingual Information Retrieval).

The ability to search and retrieve information in multiple languages is becoming increasingly important and challenging in today's environment. Consequently, multilingual and cross-lingual (language) information retrieval (MLIR and CLIR) search engines have received more research attention and are increasingly being used to retrieve information on the Internet. Cross-lingual IR has become more important in recent years. CLIR refers to searching, translating and retrieving information in different languages, but mainly between a source language and a target language.

The paper presents a detailed survey of various CLIR techniques and advantages and limitation of each method.. CLIR techniques classification based on the research fields and their comparison is discussed.We cite the past literature,types of CLIR techniques based on research fields are discussed and finally the conclusion.

Cross lingual information retrieval

Cross-language information retrieval (CLIR) is a retrieval process in which the user presents queries in one language to retrieve information in another language. CLIR approaches are decomposed into two research fields :- the first is dictionary-based approach[bilingual MRD and machine translation (MT)], and the second is concept driven approach.

In dictionary based query translation the query keywords are translated to the target language using Machine Readable Dictionaries (MRD). MRDs are electronic versions of printed dictionaries, and may be general dictionaries or specific domain dictionaries or a combination of both. The major problem in the bilingual dictionary approach is translation ambiguity in addition to problems of word inflection, problems of

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

translating word compounds, phrases, proper names, spelling variants and special terms. MT systems normally attempt to determine the correct word sense for translation by using context analysis.

Concept driven approaches such as thesauri and multilingual ontologies bridge the gap between the linguistic term and its meaning. A Bilingual Thesaurus groups words with similar meanings in hierarchies (with several levels) of classes and sections and maps them according to their meanings. However, the thesaurus does not include the definition of words. Ontology is a generalized collection of knowledge that will be used to add a context to search queries by the query expansion, enabling word sense disambiguation. So the paper focuses CLIR approach using ontology rather than collecting a thesaurus.

Literature survey

Mustafa abusalah et al[1] reports an experiment to evaluate a Cross Language Information Retrieval (CLIR) system that uses an ontology to improve query translation in the travel domain. The ontology-based approach significantly outperformed the Machine Readable Dictionary translation baseline using Mean Average Precision as a metric in a user-centered experiment.

Fedric.C.Gey and nine researchers[2]had proposed a method that can potentially acquire all the parallel texts in a web site using cross-lingual information retrieval (CLIR) techniques using parallel corpora based technique.

Ari Pirkola, Turid Hedlund et.al[3] reviewed literature on dictionary-based cross-language information retrieval (CLIR) and presents CLIR research done at the University of Tampere (UTA). The structured query model and report findings for four different language pairs concerning the effectiveness of query structuring is presented. Pattabhi R.K Rao and Sobha. L[4] have described how cross lingual information retrieval can be effectively done between a highly agglutinative language, Tamil and English, an isolating language. The query needs to be processed using a morphological analyzer or a stemmer to obtain the base forms of the given query terms.

Feng YuI, Dequan Zheng et.al[5]have showed that for improving the effectiveness of cross-lingual information retrieval (CLIR), a domain ontology knowledge based method is presented to apply to C-E CLIR. In this study, the domain ontology knowledge is acquired from both source language user queries and target documents to select target translation and re-rank initial retrieval documents set. Manoj Kumar Chinnakotla et.al[6] have discussed a query translation based approach using bi-lingual dictionaries. Query words not found in the dictionary are transliterated using a simple rule based transliteration approach. Using the above approach, for Hindi, a Mean Average Precision (MAP) of 0.2366 using title and a MAP of 0.2952 using title and description are achieved. For Marathi, a MAP of 0.2163 using title is achieved.

Sujoy Das et.al[7] made an observation that the dictionary based query translation approach has been widely used by researchers of CLIR. The translation ambiguity and target polysemy are the two major problems of dictionary based CLIR. In this paper, the researchers have investigated part of speech and co-occurrence based disambiguation techniques for English-Hindi CLIR system. Dinesh Mavaluru and Dr. R. Shriram [8] had proposed a hybrid model for Telugu English CLIR system. The bi-lingual ontology is used to convert the Telugu word to the English word. The overall system has implemented in Java and the Ontology has built for Telugu language. The system has tested for accuracy. The Google search interface was used.

Saurabh Varshney and Jyoti Bajpai [9] have studied the effect of target polysemy and translation ambiguity in dictionary based query translation approach for English- Hindi CLIR system.Nurjannaton Hidayah Rais et.al[10] discusses research on query translation events in Malay-English Cross-Language Information Retrieval (CLIR) system. The researchers have assumed that by improving query translation accuracy, they can improve the information retrieval performance.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Debasis Mandal et.al[11] describes the experiment on two cross-lingual and one monolingual English text retrievals. A careful analysis of the queries revealed that the queries with named entities provided better results for all the runs, whereas the queries without named entities performed very poor due to poor bilingual lexicons and thus bringing down the overall performance metrics. S.M.Chaware and Srikanth Rao [12] discusses an approach to build ontology from relational database with some additional rules. The ontology can be build dynamically as per user's need, which will give overall knowledge domain to the user. The result shows the complete, easy and simple way of building ontology from database.

Mustafa Abusalah et.al[13] have reviewed a literature survey based on CLIR system in short. They have described previous work in CLIR, current problems in CLIR, and made recommendations for future work.

Types of CLIR techniques

There is an increasing amount of full text material in various languages available through the Internet and other information suppliers. Therefore cross-language information retrieval (CLIR) has become an important new research area. Some CLIR systems use language resources such as bilingual dictionaries to translate the user's original query, while other systems use machine translation to translate the foreign-language documents beforehand, enabling them to be retrieved by the original query.

Based on the first research field which is "Bilingual MRD and MT",CLIR system is divided into:-

1] Query Translation
2] Document Translation

Based on second research field which is "Concept based field", CLIR system has two approaches:-

1] Multilingual dictionary (Dictionary based).
2] Ontology

CLIR techniques can be classified into different categories based on translation resources:

1] Dictionary-based CLIR technique (DB-CLIR)
2] Corpora based CLIR technique
3] Machine translator based CLIR technique
4]Ontology based CLIR technique (Concept driven field)

Dictionary-based CLIR technique

In dictionary based query translation the query keywords are translated to the target language using Machine Readable Dictionaries (MRD) abbreviations in the title or heads unless they are unavoidable. DB-CLIR is applied using document translation as well as query translation

DB-CLIR using document translation

There are two main strategies in DB-CLIR:-

i) translating the original documents into the language of the queries,

ii) translating the queries into the language of the documents.

This family of approaches includes all techniques which rely on a simple machine-readable bilingual dictionary to map the bag of words query derived from the user request to a semantically equivalent bag of words representation in the document language.

Using the dictionary based translation is a traditional approach in cross-lingual IR systems but significant performance degradation is observed when queries contain words or phrases that do not appear in the dictionary.

DB-CLIR using query translation

The document translation approach requires that the entire documents in the collection are translated into the language of the user request. The approach may require enormous translation effort and will be expensive. In query translation approach the query is translated into the documents language and then monolingual retrieval is performed. The query can be translated using machine translation system, parallel texts and/or domain specific corpora, or Machine Readable Dictionary MRD. Query translation approach is popular among CLIR community because it is efficient and easily implemented for relatively short queries.

Corpora based CLIR technique

A Corpus is a repository of a collection of natural language material, such as text, paragraphs, and sentences from one ormany languages. Two types of corpora (plural of "corpus") have been used in query translation:-

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

- Parallel corpora
- Comparable corpora

Parallel corpora

Parallel corpora consist of the same text in more than one language. When retrieving text from a parallel corpus, the query in this does not need to be translated, since a source language query can be matched against the source language component of the corpus, and then the target language component aligned to it can be easily retrieved.

Comparable Corpora

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 number of statistical techniques can be used to

derive topic-specific (often technical) bilingual dictionaries from parallel corpora.

The corpora based CLIR technique mainly consists of four modules: preprocessing, candidate texts retrieval, parallel texts verification, and duplicate elimination.

i) At first, web pages are inputted into preprocessing module, texts in each page are extracted and saved with the page's URL and the texts' relative location in the page.

ii) Candidate texts retrieval module then builds index for English texts, and retrieves those English texts mostly similar to any language text by a cross-lingual information retrieval model, the retrieval results will contain the wanted parallel texts.

iii) In the next step, parallel texts verification module selects those real parallel texts from candidate texts. Finally, duplicate elimination module checks duplicate texts and removes excess ones.

Results returned by this type of cross-lingual information retrieval technique contain a considerable amount of parallel texts with over 90% precision

Machine Translator based CLIR technique

In CLIR, Machine Translation (MT) can be implemented in two different ways. The first way is to use an MT system to

translate foreign language documents in the corpora into the language of the user's query. This is done off-line beforehand. This approach is not viable for large document collections, or for collections in which the documents are in numerous languages.

In the second method of using MT in CLIR, the users query in the "source" language is translated into the "target" language (the language of the documents in the stored collection). The "target" language query is then used to retrieve "target" language documents using classical IR techniques. With both methods, the MT stage is separate from the retrieval stage. An ambiguity problem exists in the MT component, since the translated query does not necessarily represents the sense of the original query. For instance, translating the English query big bank to another language could produce an inappropriate translation since it is not clear whether "bank" means the institution or the edge of a river. MT systems normally attempt to determine the correct word sense for translation by using context analysis. However, a typical search engine query lacks context as it consists of a small number of keywords. MT is more efficient in documents translation as the context is clearer

Compared with dictionary or corpus based methods, the advantage of MT-based CLIR translation lies in that technologies integrated in MT systems, such as syntactic and semantic analysis, could help to improve the translation accuracy

Ontology based CLIR technique

This type of technique comes under the field of Concept driven research. Concept driven approaches such as thesaurus (multilingual dictionary) and multilingual ontologies bridge the gap between the linguistic term and its meaning. Ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base. An ontology, together with a set of instances of the classes or concepts defined, constitutes a

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

knowledge base about the domain being described. The technique is as follows:-

i) The Ontology based IR system submits the query keywords to XSL (Extensible style sheet Language) to query the ontologies, extracting related concepts and concept relations.

ii) Then concepts associated with semantic relations are studied by the ontology based CLIR system .

iii) They are then identified for query expansion if synonyms were found, this is all done monolingual, then concepts are translated into their equivalent concepts in the other language using the ontology bilingual index.

iv) If the concept was not found in the ontology, the Dictionary is used to find the relevant translated concepts.

v) The final translated query terms are combined using the Boolean OR and then matched with the corpora documents.

vi) The results then are ranked depending on many factors such as the number of matching terms found in each document and the number of terms occurring in the document.

In this technique, the ontology is represented in XML form so that concepts, sub-concepts and its relationships can be mapped easily. Being able to identify the most appropriate translation results of ontology concepts is crucial in the ontology translation phase.

Simple approaches have been developed for CLIR by using multi-lingual dictionary or Word Net. Ontology will be better choice for CLIR, as it covers the entire context and its relationships, which will be helpful for both user and system provider. To acquire knowledge, even if for a small activity, every time there is need to access entire database. It will decrease the performance in terms result and also time consuming. In order to avoid this activity and to improve the performance, ontology is the best solution. Whenever any knowledge is required, data about that sub-domain can be considered, an ontology gives the accurate knowledge.

Ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base. According to this definition, the same ontology can be used for building several KBs, which would share the same skeleton. These skeletons can be extended by adding concepts and sub-concepts that cover new areas. Such ontology will give easy and clear understanding of structure of ontology and inference mechanisms will become easier.

Following table presents a comparison of all CLIR techniques based on the research fields.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Table 1: Comparison of CLIR techniques

Types of CLIR echnique	**Subtypes**	**Concept**	**Advantage**	**Disadvantage**	**Application**
DB-CLIR	Using Document translation	The document translation approach requires that the entire documents in the collection are translated into the language of the user request.	It relies on a simple machine-readable bilingual dictionary to map the bag of words query derived from the user request to a semantically equivalent bag of words representation in the document language.	(1)untranslatable search keys due to the limitations of general dictionaries, (2) the processing of inflected words, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages	Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings [1]
	Using Query translation	In the query translation approach, parts of speech of a word in the given context is found and the process of translation and transliteration is performed.	This approach overcomes translation ambiguity and target polysemy which are the major drawbacks of DB-CLIR using document translation.	The queries with named entities provided better results whereas the queries without named entities performed very poor due to poor bilingual lexicon	1)Hindi to English and Marathi to English CLIR [3] 2)Query Translation Architecture for Malay-English CLIR [6]
Types of CLIR echnique	**Subtypes**	**Concept**	**Advantage**	**Disadvantage**	**Application**
Corpora based	-	Each text in source language is associated with the texts potentially parallel to it in a CLIR process and then each text pair is verified by a pattern based algorithm.	The estimation of translation matrix generated during the transformation of query language can be done efficiently by reducing the document space. The results of these estimations are promising.	Corpora based methods suffers lack of resources. Parallel corpora are not always readily available and those that are available tend to be relatively small or to cover only a small number of subjects.	The TEC-2001: Cross Language Information Retrieval Track [2]

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

MT-based	Using Query translation	A query translation is conducted with the degraded MT systems and translated queries of varying quality are obtained. Then the translated queries are submitted to the IR system and performance is evaluated.	Technologies integrated in MT systems, such as syntactic and semantic analysis, could help to improve the translation accuracy	An ambiguity problem exists in the MT component, since the translated query does not necessarily represents the sense of the original query.	Literature review of Cross Language Information Retrieval[13]
Ontology based	-	The Ontology based IR system submits the query keywords to XSL (Extensible style sheet Language) to query the ontologies, extracting related concepts and concept relations.	1) Since the ontology is represented in the form of XML, concepts, sub-concepts and relationships can be mapped easily. 2) Enables reuse of domain knowledge	1) Increases the creation difficulty. 2)Visualization problems The size of the resource (ontology) is inversely proportional to its specificity.	Chinese-English CLIR based on Domain Ontology Knowledge[5] Ontology Approach for Cross-Language Information Retrieval [12]

Conclusion

The Internet has paved opportunities for increasing multi-lingual information exchange and retrieval in future. Cross-lingual IR provides new paradigms in searching documents through myriad varieties of languages across the world and it can be the baseline for searching not only among two languages but also in multiple. Creating accurate metadata in different languages in documents or good translation of key information in documents can help improve the quality of the index and retrieval. After the evaluation of both the pure dictionary and the ontology systems, the ontology based system scored higher in terms of precision.

In future development ontology will be enhanced and extended by using annotation tools to align new concepts to the ontology and then test it again with the dictionary system. Other areas for investigation include ease of use, the use of relevance feedback, the effect of more extensive use of concept relations and possibly experiments with larger data sets. This paper discussed different types of CLIR techniques and advantages and disadvantages of each techniques.

References

[1] Mustafa Abusalah, John Tait and Micheal Oakes "*Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base*" September 2009.

[2] F. C. Gey, "*The TEC-2001: Cross Language Information Retrieval Track,"* 2001.

[3] Ari Pirkola, Turid Hedlund, Heikki Keskustalo,andKalervoJärvelin,"*DictionaryBased CrossLanguage Information Retrieval: Problems, Methods,and Research Finding*" September 2001, Volume 4,pp

[4] Pattabhi R.K Rao and Sobha. L **,**"*Cross Lingual Information Retrieval Track*",AU-KBC Research Centre, MIT Campus, Chennai,2010

[5] Feng YuI, Dequan Zheng and Tiejun Zhao, Sheng Li, Hao Yu, "*Chinese-English Cross-Lingual Information Retrieval based on Domain Ontology Knowledge*",2010

[6] Manoj Kumar Chinnakotla , Sagar Ranadive, Om P. Damani, and Pushpak Bhattacharyya , "Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation ", Department of Computer Science and Engineering, IIT Bombay, India,2008

[7] Sujoy Das, Anurag Seetha , M. Kumar and J. L. Rana, "Disambiguation Strategies for English-Hindi Cross

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Language Information Retrieval System",2009

[8] Dinesh Mavaluru Dr. R. Shriram, " Telugu English Cross Language Information Retrieval: A Case Study ", 2013

[9] Saurabh Varshney and Jyoti Bajpai, " Improving performance of English-Hindi cross language information retrieval using transliteration of query terms", 2013

[10] Nurjannaton Hidayah Rais,Muhamad Taufik Abdullah, Rabiah Abdul Kadir, "Query Translation Architecture for Malay-English Cross-Language Information Retrieval System",2010

[11] Debasis Mandal, Sandipan Dandapat, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar, "Bengali and Hindi to English Cross language Text Retrieval under Limited Resources",2008

[12] S.M.Chaware and Srikanth Rao, "Ontology approach for cross language Information Retrieval",2011

[13] Mustafa Abusalah, John Tait and Micheal Oakes "Literature review of Cross language information retrieval",2007