ISSN 2229-5518

Survey on Cross-Lingual Information Retrieval

[1]*Kehinde Kayode Agbele(PhD)*
[1]Department of Computer Science
Elizade University
Ilara-Mokin, Nigeria

[2]*Eniafe Festus Ayetiran*, [3]*Kehinde Daniel Aruleba*
[2,3]Department of Computer Science
Elizade University
Ilara-Mokin, Nigeria

***Abstract*—** The rise in unmatched multilingual resources afforded by the exponential WWW growth demands the advancement of technologies to eradicate the communication barriers among languages. Relevant information in collections and the Web is not limited to the native language of the user, but today, the need to retrieve documents in other languages is growing so that the content, which can be translated, satisfies the information needs of the user. Information retrieval (IR) can be classified into different categories such as monolingual information retrieval, Cross lingual information retrieval (CLIR) and Multi lingual information retrieval (MLIR). In the present day scenario, the diversity of information and language barriers are the serious challenges for communication and cultural interchange across the globe. To solve such communication barriers, CLIR systems are today in strong demand. The goal of CLIR is to find relevant information written in a language different from other languages of the query. CLIR can be used to improve the capabilities of users to search and retrieve documents in many languages. Diverse translation techniques can be used to achieve CLIR. In this paper, we review the techniques and approaches of CLIR research for query and document translation and their role in current research directions, which include new models, and paradigm in the extensive area of IR. In addition, based on existing literature, a number of challenges and tools in CLIR has been identified and discussed. Finally, possible future research directions on semantic query-document translation for CLIR are discussed.

Keywords*—CLIR; information retrieval, query translation, document translation, corporal based translation, machine translation, corporal-based translation, word sense disambiguation*

I. *INTRODUCTION*

Cross Language Information Retrieval (CLIR) is a sub-field of Information Retrieval (IR) that deals with retrieving relevant information stored in a language different from the language of user's given query. The purpose of CLIR is to provide the benefits to the user in finding and accessing information without being limited by language barriers. However, with the popularity of the internet technology and increase in available online resources (data), the demand of searching for information from multi-lingual documents is increasing at an alarming rate, which results in the great challenge of how to match the user's query written in one language with the documents written in other languages. According to [Gaillard et. al 2010], CLIR provides a suitable way that can address the problems of language boundaries, where users can submit queries written in their own language and retrieve documents in other languages [Pigur, 1979]. With the rapid advancement of internet, globalization of information structure caused the urgent demand for CLIR, because CLIR allows the usage of information interchanges between diverse languages, remove linguistic disparity between the queries that are submitted and documents that are retrieved using resources over the network, which also decreases the communication cost. [Peng et. al. 2008].

The research on IR came into existence forty-six years ago whereas experiments for retrieving information across languages were first originated were first originated by [Salton, 1973]. Nevertheless, most of the modern research on CLIR started twenty-six years ago, and today it has become one of the most vital research topics in the area of IR. An ever active research field, a huge number of researches and studies have been published on CLIR and various issues are addressed in numerous evaluation forums such as TREC [Voorhees et. al 2005] and CLEF [Gey et. al 2008] while each of them cover different languages. For example TREC covers Spanish, Chinese, German, French, Arabic and Italian and CLEF covers French, German, Italian, Spanish, Dutch, Finnish, Swedish and Russian [Ahmed and Nurnberger 2012]. The most effective way to unravel 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 approaches. Our particular emphasis in this survey is on query translation approach to translate the languages using translation techniques for CLIR.

The remainder of the article is structured as follows: **Section 2** describes the query translation, document translation and query-document translation approaches respectively; **Section 3** describes the comparative analysis of the three approaches in literature; **Section 4 & 5** describes the challenges of CLIR and CLIR tools respectively; **Section 6** describes the State-of-the art algorithm and techniques in CLIR; **Section 7** describes the several methods that can be used to evaluate the output of a translation system in the context of CLIR and **Section 8** concludes the survey with a look at the future of CLIR.

2.1 QUERY TRANSLATION APPROACH

Query translation can be based on using bilingual dictionary or using the corpora or machine translation. The key challenge in CLIR is to bridge the language gap between query and documents. The authors in [Narasimha et. al 2014; Wu and He 2010; Oard et. al 2008] reported that query translation is now serving as a major Cross-lingual mechanism in current CLIR systems. CLIR search engines enable users to retrieve content in a language different from the language used to formulate the query. Translation of query has the advantage that the computational effort (time and space is less when compared with other methods. Query translation has the following disadvantages; (i) usually a query do not provide enough contexts to automatically find the anticipated meaning of each term in the query. (ii) Translation errors affect retrieval performances sensibly. (iii) In case of searching a multi-lingual database, query has to be translated into each of the languages of the database. In CLIR, query translation play a vital role that can be achieved by the following approaches: dictionary based translation approach, corpora based translation approach and machine based translation approach respectively.

2.1.1 Dictionary-Based Translation Approach

In Dictionary-based query translation, the query is processed linguistically and only keywords are translated using the Machine Readable Dictionaries (MRD). MRDs are electronic version of printed dictionaries, either in general domain or in specific domain. The use of existing linguistic resources, especially the MRDs, is a natural approach to Cross Lingual IR. Translating the queries using the dictionaries is much faster and simpler than translating the documents according to [Carley 1999; Aljlayl et. al 2001; Pirkola et. al 2001] in [Seetha et. al 2007], the following are common problems related with dictionary-based translation:

(i) Untranslatable words (like new compound words, proper names, spelling variants and special terms): Not every form of words used in a query s always found in the dictionary. Sometimes, problem occurs in translating different compound words (formed by a combination of new words) due to the unavailability of their proper translation in the dictionary [Pfeifer et. al 1996].

(ii) Processing of inflected words: Inflected word forms are usually not found in dictionaries [Fluhr et. al 1998].

(iii) Lexical ambiguity in source and target languages: Relevant forms of lexical meaning for information retrieval are 1) homonymous and 2) Polysemous. Two words are homonymous if they have at least two different meanings and sense of words is 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 [Lyons 1981]

2.1.2 Corporal-Based Translation Approach

Query translation using corpora require 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 [Picchi & Peters 2000; Landauer & Littman 1990] queries are translated based on multilingual terms extracted from parallel or comparable documents collections. A parallel corpus has been used since the early 1990s for translation of given words. A parallel corpus is a collection of text, 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 [Chandra and Dwivedi 2014; Manning et. al 2008].

A comparable corpus is one of the important concepts in corpus-based translation study introduced by Baker [Fernandez 2006]. 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 good translations of each other [Munteanu and Marcu 2005; Landauer and Littman 1990; Sheridan and Ballerini 1996].

2.1.3 Machine Translation Approach

Cross-lingual IR with query translation using machine translation [Carley 1999] seems to be an obvious choice compared to dictionary and corpora-based approaches. The advantages of using the machine translation is that it saves time while translating large texts. Manning and Schutze [2008] distinguished four different approaches to deal with machine translation, which includes: (a) Word-for-word approach, (b) Syntactic transfer approach, (c) Semantic transfer approach, and (d) Interlingua approach. The ultimate goal of CLIR machine translation (MT) systems is to translate queries from one language to the other by using the context. Many factors contribute to the difficulties of machine translation, which includes 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 criticism mostly stem from the fact that the current translation quality from 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. MT based approach 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:

SN	**Comparison between query translation Techniques**			
	Parameter	***DBT approach***	***CBT approach***	***MBT approach***
1	Development	Less expensive	More expensive than DBT	More expensive than both DBT & CBT
2	Ambiguity	High	Low	Low
3	Translation availability	Highly available in many languages	Available only in few languages	Available only in few languages.
4	Offline Translation	Likely	Likely	Not Likely

2.2 DOCUMENT TRANSLATION APPROACH

In [Croft et. al 1991; Buckley et. al 1995], document translation can be the most desirable scenario in CLIR, if the motivation is to allow the users to search the documents from their own language and receive results back in user's language. Towards this, 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 [Ramanathan 2003]. The function of this translation is in 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. However, 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 the user can understand. 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 a different group 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 [Braschler and Schauble, 2001; Franz et. al 2000; Oard and Hackett, 1998] since more varied context within each document is available for translation which can improve translation quality. The document translation approach has certain benefits over query translation. This includes 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) Translation 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 run time.

However, there are certain negative issues with document translation as well. This includes:

(a) Much more computational effort is needed to index collection

(b) Bad scaling performed in case of more than two languages

2.3 DUAL TRANSLATION-QUERY AND DOCUMENT TRANSLATION APPROACH

In this approach, both queries and documents are translated into a common representation. This approach requires additional storage space translated documents but provides scalability when same collection of documents is require in multiple languages. One of the examples of such approach is controlled vocabulary systems [Oard and Dorr 1996]. These systems represent all documents using a predefined 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 require interfaces to the vocabulary in order to be a able to generate effective queries. Dual translation approach is also called a hybrid translation approach and can be performed by pivot language. Direct translation between two languages may not always be possible due to the limitation of translation sources. To perform such type of translation, a resource 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 document is translated first into a pivot language, then to the target language; translate both document and query into pivot language.

3. COMPARATIVE ANALYSIS OF THE THREE EXISTING APPROACHES

The need for translation has itself been questioned because non-translation based methods of CLIR such as cognate-matching [Buckley et. al 1998] and cross-language latent semantic indexing [Dumais et. al 1997] have been developed. Document translation into query language or query translation into document language is the two approaches that coupled machine translation and information retrieval. Query translation and document translation 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 researchers suggest that document translation should be competitive or superior to query translation. Typical queries are short and may contain keywords or phrases only when these are translated inappropriately, the IR engine has no chance to recover. In translating a long document, MT engine offers many more opportunities to translate keywords 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 interaction 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 is to 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 document 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 other languages. Document translation can be performed offline and online but query translation is performed only online. Hybrid system that uses both query and document translation are possible because of the trade-off between computer resources and the 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 match unambiguously. In this approach, the additional storage space requirement is independent of the number of languages supported. The major problems that occur 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 document translation) of CLIR are described in Table 2 while table 3 describes the comparative analysis of the three approaches of CLIR

TABLE 2:

SN	**Difference between Query and Document translation Approaches**		
	Parameter	***Query Translation approach***	***Document Translation approach***
1	Language	Previous knowledge of translation is not required.	Previous knowledge of translation is required.
2	Ambiguity	Maximum chances of occurring ambiguity	Minimum chances of occurring ambiguity
3	Size	Small	Large
4	**Recovery**	When these are translated inappropriately, the IR engine has no chance to recover	Chances to recover exist.
5	Overhead	Low	High
6	Cost	Low	High cost

TABLE 3:

SN	**Comparison of the Three Translation Approaches**			
	Parameter	***Query Translation***	**Document Translation**	**Query-Document Translation**
1	Extra storage space	Not needed	Needed	Not needed
2	Ambiguity	Maximum	Minimum	More than both query and document
3	Information Retrieval	Bilingual	Bilingual	Bilingual and Multilingual
4	Transition Time	Less	More than query	More than both query and document
5	Flexibility	Highly	Less	Less

4. CHALLENGES IN CROSS-LINGUAL IR

Each of the approaches in sections 2.1, 2.2 and 2.3 has created challenges to the CLIR. One of the problems is the translation disambiguation. 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 divergence 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 overcome the barrier between the source language (SL) in query sentences and the target language (TL) in documents to be searched. As discussed in sub-2.1.3, 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. The challenges of CLIR are discussed in the following sub-section.

4.1 Ambiguity

In (Chandra and Dwivedi, 2014), ambiguity occurs when words have multiple meaning, which also referred to as homonymy or polysemy. Ambiguities in IR are semantic and synthetic in nature, whereas ambiguities in CLIR are semantic and lexical. Therefore, the probability of occurrence of ambiguity in CLIR is higher than normal IR, due to the availability of different languages (Diekema and Anne 2003).

4.2 Knowledgeable Terminology

Knowledgeable terminology, for example scientific names, is often problematic and is often found in knowledgeable dictionaries or term banks. Knowledgeable terminology tends to be less ambiguous than regular vocabulary while regular vocabulary can have a knowledgeable meaning when used in a certain subject domain.

4.3 Effective User Feedback

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

4.4 Difficulty in State-of-the-Art Applications

Question and answering is relatively a new stream of information retrieval. In question and answering, end-users throw a question in a form of query and retrieval answers for that in order to satisfy the user information needs. However, the major challenge is to retrieve answers of English questions in a different language other than the native language of the user.

5. CROSS-LINGUAL IR TOOLS

In this section, we describe the existing CLIR systems. Over the past few years, research in CLIR has progressed and many systems have been developed. Some prominent systems of CLIR are described in the following:

5.1 KANSHIN- KANSHIN collects and analyses the multilingual articles of Japanese, Chinese, Korean and English languages (Fukuhara et al. 2008). The system provides a various viewpoints for user such as temporal, focal, geographical, and network. It also provides a cross-lingual keyword navigation tool between slog survey tools (called splogExplorer) and inters language links of Wikipedia.

5.2 KEIZAI- KEIZAI (Ogden et al. 1999) 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 retrieval documents together with small images, which are called Document Thumbnail Visualization. The advantage of visualization is to improve the recall and efficiency.

5.3 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 (He et al. 2003).

5.4 MULINEX - MULINEX system (De Luca et al. 2006) 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 categorization algorithms used in Mulinex for different tasks: n-gram categorizer for noisy input, k-nearest-neighbor (KNN) algorithm for normal documents and pattern categorizer for every short document.

5.5 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 (Hersh et al. 1998). It provides an intelligent healthcare monitoring architecture for high quality health care services with reasonable cost.

5.6 UCLIP - 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 (Abdelali et al 2004).

6. STATE-OF-THE-ART ON CROSS-LINGUAL IR

Wikipedia has become an important resource in the cross-lingual IR recently. Many researchers have conducted studies and experiments using the free online encyclopedia. In [Lin et al 2009], the authors developed a Japanese-Chinese IR system based on the query translation approach. The system employed a more conventional Japanese bilingual dictionary and Wikipedia for translating query terms. They studied the effect of using Wikipedia and proposed that Wikipedia can be used as a good Named entities (NEs) bilingual dictionary. To cope with term disambiguation, the authors have adopted an iterative disambiguating method based on the PageRank algorithm. The method proved effective and outer performed the previous Japanese-Chinese systems tests.

A recent Wikipedia-based study by [Nguyen et al. 2009] showed that query translations for cross-lingual IR can be performed using only Wikipedia. An advantage of using Wikipedia is that it allows translating phrases and proper nouns wells. It is also very scalable since it is easy to use the most up-to-date version of Wikipedia, which makes it able to handle actual terms. The approach is that the queries are mapped to Wikipedia concepts and the corresponding translations of these concepts in the target language are used to create the final query. Wiki Translate system [Nguyan et al. 2009] is evaluated by searching the topics in Dutch, French, and Spanish language within an English data collection. The system, which achieved a performance of 67% compared to the monolingual baseline, can be a valuable alternative to current translation resources. The unique structure of Wikipedia (for example the text and internal links) can be very useful in cross-lingual IR. The use of Wikipedia might also be suitable for interactive CLIR, where user feedbacks are also used to translate the query, since Wikipedia is already very popular among internet users.

Query suggestions aim to suggest relevant queries for a given query, which help users to specify their information needs better [Gao, W., et al 2007]. It is closely related to query expansion but query suggestions will suggest full queries that have been formulated by users in another language. Gao et al [2007] proposed query suggestions by mining relevant queries in different languages from up-to-date query logs as it is expected that for most user queries, we can find common formulations on these topics in the query log in the target language. Therefore, cross-lingual query suggestions also play a role of adapting the original query formulation to the common formulations of similar topics in the target language. Used as a query translation system, the proposed method demonstrations higher effectiveness than traditional translation methods using bilingual either dictionary or machine translation tools.

Pourmahmod and Shamsfard in [2008] carried out a research to retrieve English documents relevant to Persian queries using bilingual ontologies to annotate the documents and queries. A bilingual ontology consists of ontology and a bilingual dictionary. Ontology is a formal, explicit specification of a shared conceptualization. It contains a set and identified concepts related by a set of relations [Shamsfard et al]. They used the ontology to expand the query with related terms in pre—and post-translation expansion and the combined approach significant improves cross-lingual performance. Researchers in [Lilleng and Tomassen, 2007] analyzed the query translation in cross lingual IR based on feature vectors and usage of context information during the query translation. They pointed out that by using information external to the query, such as the ontologies and document collections, the effects of disambiguation and polysemy can be reduced. The characteristics of a feature vector are dependent on the quality of both the ontology and the document collection being used. As the research is still in progress, they still need to fully implement the approach for more thorough testing and evaluation. However, an advantage of this approach is the adaptability to several languages, which can be done by adding other dictionaries and thesauruses.

Disambiguation is the aim of most translation techniques used in CLIR. Yuan and Yu [2007] found a method using co-occurrences between pairs of terms as statistical measure, unlike the traditional statistical approach. This method needs only a bilingual dictionary and a monolingual corpus for translation. They compared different combinations of target terms and presented the output in the form of probability distribution. Using the results, the query is converted to target language. It is a simple method and experiment showed that it performed well.

The increasing numbers of multi-lingual documents in web posed a challenge in managing them. Wu and Lu [2007] identifies novel model called domain alignment translation model to conduct cross-lingual document clustering and term translation simultaneously an in the end the multi-lingual documents with similar topics can be clustered together. Their method with the use of only a bilingual dictionary can achieve comparable performance with the machine translation method using Google translation tool. Although their experiments only consider word but ignoring the base phrase, the clustering in the source language and the clustering in the target language are related highly and the clustering quality can be emphasized for future research.

7. EVALUATION STUDY FROM CLEF AND TREC

In this section, we describe some common approaches that can be used when evaluating the quality of a translation system in the context of cross-language information. We discuss evaluation at this specific point in the paper to provide a common vocabulary for the comparison of various techniques and translation models in the following sections.

Evaluating the effectiveness of a query and/or document translation usually involves assessing the retrieval effectiveness of the CLIR engine associated with it. The standard mechanisms for this sort of assessment all rely upon the availability of large Cranfield-style test collections [Cleverdon 1991]. A Cranfield-style test collection normally consists of a document corpus, a set of search topics, and a matched set of assessments. The document corpus is often provided as a set of semi-structured XML documents. Each document in this set will consist of several text fields {e.g., title, abstract, keywords} and a unique identification number. The search topics will describe a number of search tasks. These tasks are often categorized as short tasks or long query tasks. The long query tasks are more detailed, but short tasks tend to replicate realistic web queries.

Relevance assessments are manually derived assessment representing the relevance to each topic of each document in the corpus. Producing these assessments is a time-consuming and expensive process. Because test collections are generally large, only a fraction of the documents relevant to each query is scored. The standard approach to selecting this subset is known as pooling [Kuriyama et al. 2002]. A pooling operation selection only the k documents returned by a number of different CLIR systems for manual assessment. The CLIR engines used during pooling run are usually the same CLIR engines that need to be evaluated. Monolingual runs interactive runs (either cross-lingual or monolingual) can also be used to enrich the pools.

The production of test collections has always been one that main responsibilities of CLIR conferences and workshops. The Text Retrieval Conferences (TREC), sponsored by the National Institute of Standards and Technology (NIST), was started in 1992 as part of the TIPSTER Test program. Its purpose was to support research within the IR community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. TREC had a very important influence on CLIR, especially in its formative years, and hosted one of the earliest competitive CLIR tracks in 1997 (TREC-6). DARPA was another early proponent of CLIR, launching the TIDES (Translingual Information Detection, Extraction, and Summarization) program. Other CLIR conferences of note include the Cross-Language Form (CLEF), which concentrates on European languages, and the Form for Information Retrieval Evaluation (FIRE), a CLEF-2007 spin-off chiefly concerned with Indian language (Hindi, Telugu, and Malayalam). Along a similar line, (NTCIR) was founded, in 1999. It has been responsible for a series of evaluation workshops designed to enhance CLIR research in Pacific Rim language (e.g Chinese, Japanese, and Korean).

Researchers with access to a test collection can retrieve documents using topics provided and then measure that are applied at this stage are precision, recall, and the relevance judgements. The three basic measures that are applied at this stage are precision, recall, and the F-measure (also known as the F1 score). Precision is the fraction of retrieved documents that are relevant. Recall is the fraction of relevant documents that are retrieved. The F-Measure is the harmonic mean of precision and recall. These are set-based measures that are computed using unordered sets of documents. When working with ranked retrieval results, a number of other measurements may come into play [Manning et al. 2008]. In these circumstances, precision can be measured suing a relatively low number of retrieved results (e.g., precision 10 results). Mean average precision (MAP) can provide a single figure measure of quality across recall levels. Normalized discounted cumulative gain (NDCG) can be deployed in situations involving non-binary notions of relevance. There are many other measurements that can apply [Baeza-Yates and Ribeiro-Neto 2008]. Once measured, the retrieval effectiveness of a CLIR system is often compared with a monolingual baseline. Tests of statistical significance, such as the widely used Wilcoxon signed-rank test, are commonly used when interpreting results [Hull 1993].

8. FUTURE WORK

Due to the problems inherent in dictionary-based, corpus-based and machine translation approaches, we propose a sense-based approach, which uses multi-lingual lexical resources such as BabelNet to computationally determine the sense of the word to be translated. This will go a long way to solve the problems discussed earlier.

9. CONCLUSION

In this survey, we have outlined the various types of techniques that can be used when translating queries and/or documents in the context of CLIR. Cross-lingual IR provides new paradigms for searching documents through countless diversity of languages across the globe and it can be the baseline for searching not only between two languages but also in multiple languages. Today, most of the cross lingual involved only few famous languages like English, Hindi, Spanish, China, Japanese and French. Research on languages has increases the development of country. As the world becomes more connected by technology, CLIR in every language is needed. Cross language, information retrieval systems offer a reasonable, technically feasible mechanism through which access can be provided. CLIR is a multidiscipline area that has been increasingly ginning more attention from the research community. Despite recent advances and new developments, there are still many aspects to be explored. The purpose of this paper is to review some of the latest researches in the area of CLIR. The survey indicates that query translation is always the choice as compared to document translation. It is more convenient to translate only the query than the whole documents. Documents translation, which uses machine translation, is computationally expensive and the size of document collection is large. However, it might be practical in the future when the computer technology improves. In this paper, we explain a description on CLIR, its challenges and current methods and techniques, and future research goals to overcome problems for efficient and resourceful searching. In reviewing this information, it becomes possible to gain a larger picture the CLIR field.

Looking to the future, we anticipate a steady increase in the quality and quantity of translation resources available to researchers. There are currently no researches conducted on cross-lingual IR with emphasis on sense-based query-documents translation for cross lingual IR. It is hoped that more researchers that focus on sense-based query-document translation will be conducted in the future.

Acknowledgment

We appreciate and thank Elizade University Management for the financial assistance in making this article a reality.

References

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

[2] V. A. Pigur, (1979), “Multilanguage Information-retrieval systems: Integration levels & language support”, Automatic Documentation and Mathematical Linguistics, Vol. 13, No 1, pp. 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] G. Salton, (1973), “Experiments in multi-lingual information retrieval”, Information Processing Letters, Vol. 2, No 1, pp. 6-11

[5] T. Voorhees, M.Ellen and K. Donna. Harman, Eds, (2005), “TREC Experiment and evaluation in information retrieval. Vol 63. Cambridge: MIT press, Vol 63

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

[7] F. Ahmed, and A. Nurnberger, (2012), “Literature review of interactive cross language information retrieval tools”, int. Arab J. Info Technol., Vol 9, No 5, pp. 479-486

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

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

[10] D.W. Oard, D. He, and J. Wang, (2008). User-assisted query translation for interactive cross-language information retrieval. Information Processing & Management, Vol. 44, No 1, pp 181-211

[11] M. Buckley, J. Mitra, Wals, and C. Cardie, (1998), “Using clustering and super concepts within SMART: TREC-6”, In E.M. Voorhees and D.K. Harman, editors, The 6th Text Retrieval Conference (TREC-6).

[12] S.T. Dumais, T.A. Letsche, M.L. Littman, & T. K. Landauer, (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

[13] S.J. Mc. Carley Mc (1999, June), “Should we translate or 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

[14] M. Aljlayl, and O. Frieder, (2001, October), “Effective Arabic-English cross-language information retrieval via machine-readable dictionaries and machine translation”, In proceeding of the tenth international conference on Information and knowledge management ACM, pp. 295-302.

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

[16] A. Pirkola, T. Hedlund, H. Keskustalo, and K. Jarvelin, (2001), “Dictionary-Based Cross-Language Information Retrieval: Problems, and Research Findings, “ Information Retrieval, Vol. 4, No 3-4, pp. 209-230

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

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

[19] J. Lyons, (1981), “Language and Linguistics: An introduction”, Cambridge University Press.

[20] E. Picchi, and C. Peters, (2000) “ Cross-language information retrieval: a system for comparable corpus querying, in Cross-Language Retrieval, “G. Grefenstette, Editor. 2000, Kluwer Academic Publishing: Massachusetts. Pp. 81-90.

[21] T. K. Landauer and M.L.Littman, (1990) “Full Automatic Cross Language Document Retrieval using Lantent Semantic Indexing”, In Proc. Of the 6th Conference of UW Center for New OED and Text Research, pp. 31-38.

[22] C. D. Manning, P. Raghavan, and H. Schutze, (2008). Introduction to information retrieval (Vol 1, p. 496). Cambridge University Press.

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

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

[25] Sheridan, P., and 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.

[26] Chandra, G., Dwivedi, S. K., "A Literature Survey on various Approaches of word Sense Disambiguation (2014)," in Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium on, vol. 5, no. 4, pp. 106-109, 7-8 Dec. 2014

[27] Fernandez, Lincoln Paulo. (2006), " Corpora in Translation Studies: revisiting Baker's tipology". Fragmentos: Revista de Lingua e Literature Estrangeiras, 30.

[28] Croft, W. B., Turtle, H. R., and 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.

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

[30] Ramanathan, A. (2003). State of the Art in Cross-Lingual Information Retrieval. VIVEK-BOMBAY- vol. 15, No. 2, pp. 16-22

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

[32] Franz, M. Scott McCarley, J., and 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 Standard and Technology. Available: http://trec.nist.gov/pubs/.

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

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

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

[36] Fukuhara, T., Kimura, A., Arai, Y., Yoshinaka, T., Masuda, H., Utsuro, T., & Nakagawa, H., (2008, April) " KANSHI: 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.

[37] 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

[38] He, D., Oard, D. W., J., Luo, J., Demner-Fushman, D., Darwish, K., & Leuski, A., (2003), "Making miracles: Interactive translingual search for sebuanoand hindi". ACM Transactions on Asian Language Information Processing (TALIP), Vol. 2, No.3, pp 219-244

[39] De Luca Ernesto William, Ernesto, Stefan Hauke De Luca, Andreas Nurnberger, and Sefan Schlechtweg. (2006) "MultiLexExploer-Combining Multilingual web search with Multilingual Lexical Resources". In Proceedings of the combined workshop on Language enabled educational technology and evaluation of Robust Spoken Dialogue Systems, Germany, pp. 17-21

[40] Hersh, William R., and C. Larry Donohoe, (1998), "SAPHIRE International: a tool for cross-language Information retrieval". Proceedings of the AMIA Symposium. American medical Informatics Associations, pp. 673.

[41] A. Abdelali, J. R. Cowie, D. Farwell, D., & Ogden, W. C., (2004) "UCLIR: a Multilingual Information Retrieval Tool", Intelligencia Artificial, Revista Iberoamericana de Intelligencia Artificial, 8(22), pp. 103-110

[42] D. Nguyen, et al. (2009), WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia, in Evaluating Systems for Multilingual and Multimodal Information Access. 2009. p. 58-65.

[43] W. Gao, et al. (2007). Cross-lingual query suggestion using query lingual logs of different languages. In Proceedings of the 30th Annual International ACM SIGIR Conference on research and Development in Information Retrieval, SIGIR'07. 2007: ACM Press.

[44] S. Pourmahmoud, and M. Shamsfard (2008). Semantic Cross Lingual Information Retrieval. In 2008 23rd International Symposium on Computer and Information Sciences, ISCIS 2008.

[45] M. Shamsfard, A. Nematzadeh, and S. Motiee (2006). ORank: an ontology based system for ranking documents. International Journal of Computer Sceince, 2006. Vol. 1 No.3 pp. 225-231.

[46] S. Yuan, and S. Yu, (2007). A new method for cross-language information retrieval by summing weights of graphs, in Fourth International Conference on Fuzzy Systems and Knowledge Discovery, J. Lei Editor. IEEE Computer Society. Pp. 326-330

[47] Wu, K. and B. Lu, (2007). A refinement framework for cross language text categorization, in Springer Lecture Notes in Computer Sceince, H. Li, Editor. 2007, Springer-Verlag: Berlin Heidelberg. Pp. 401-411.

[48] J. Lilleng, and S. L. Tomassen (2007). Cross-Lingual Information Retreival by feature vectors, in lecture Notes in Computer Science (including subseries lecture in Artificial Inetlligence and Lecture Notes on Bioinformatics). Pp. 229-239

[49] C. W. Cleverdon, (1991). The significant of the Cranfield tests on index languages. In Proceedings of the 14th Annual International ACM SIGIR Conference on research and Development in Information Retrieval. ACM, New York, 3-12

[50] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi, (2002). Pooling for a large-scale text collection: An analysis of the search results from the first NTCIR workshop. Inf. Retr. Vol.5, pp. 41-59

[51] C. D. Manning, P. Raghavan, and H. Schtze, (2008). Introduction to Information Retrieval. Cambridge University Press.

[52] R. Baeza-Yates, and B. Ribeiro-Neto, (2008). Modern Information Retrieval, 2nd ed. Addison-Wesley Publishing Company.

[53] Hull, D. (1993). Using Statistical testing in the evaluation of retrieval experiments. In Proceedings of the 16th Annual International ACM SIGIR Conference on research and Development in Information Retrieval. ACM, New York, pp. 329-338.