Challenges In Cross Language Information Retrieval

Manpreet Singh Lehal

Assistant Prof. Dept. Computer Science, Lyallpur Khalsa College, Jalandhar

Abstract— Internet is a huge information resource and widely accessed by users all throughout the world. The ubiquity of online sites need to surpass the limits of language as people from every corner of the earth use them. In the contemporary times, information is available diversely and the barriers of language create hindrances in the way of effective communication across the cultures. Cross language information retrieval (CLIR) system, comes as a rescue and tries to solve the communication needs. CLIR refers to the access of information in which the user types a query in the native language and gets an output in some different language. The aim of CLIR is not to give translation but relevant information which poses major challenges. This paper takes an overview of the number of challenges and issues in CLIR like translation ambiguity, phrase identification, translation, transliteration errors, morphological analysis, OOV words.

Keywords-CLIR; Langauage;Information; Challenges; machine translation.

I. INTRODUCTION

Cross Language Information Retrieval (CLIR) is a system to obtain information from the internet sites in varied languages. The user gives a query in the native language and it returns the results or documents that are written in foreign languages. On the surface level it appears simply to be a case of machine translation where the system finds the translation of the query and retrieves information from other languages. It involves normalization to match stored indexes, and identifying how words to be weighed in a query. However Cross Language Information Retrieval [7] is different and may be easier and harder than Machine Translation. Machine translation is easy as compared to CLIR because it need to choose from a single translation for each term, and the output need to be syntactically correct. Cross Language Information Retrieval systems, use bag of words which need not be in order. A CLIR system producing results in a diverse language from a native language query thus can give multiple translation alternatives. On the other side, CLIR is harder than MT because MT systems [8] are designed for restricted domains. But Information retrieval [5] works on domain independent techniques, employing techniques that are meant to be applicable to any text type or subject. Cross Language Information Retrieval methods have to work upon all domains and thus deal with a large number of words.

CLIR is gaining importance in the research area. Many workshops are being organized to address various issues in it. Each workshop is concerned with different languages apart from English. TREC discusses the use of Spanish, Chinese, German, French, Italian, and Arabic. CLEF focusses on French, German, Italian, Swedish, Spanish, Dutch, Finnish, and Russian so far whereas NTCIR deals with Japanese, Chinese and Korean.

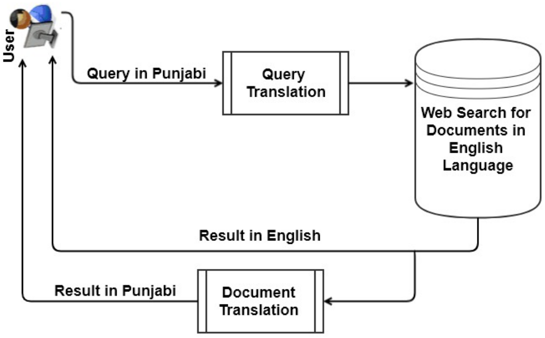

Fig.1: Cross Language Information Retrieval System

II. LITERATURE REVIEW

The research on information retrieval started as early as 1970s. The pace of research on CLIR expedited in the 90's, and it has become one of the most important research topic in the field of information retrieval. McCarley [12] compared the results of three approaches to MT based systems and a monolingual Information Retrieval system and examined their efficiencies. The MT based systems uses translation of documents and queries. The probability that a document is right answer to a query is calculated with normalization of query and document translation. Oard [13] worked further and studied the efficacy of query translation and document translation. The results may have shown higher precision for document translation than query translation however monolingual retrieval still remains on the upper edge. Fujii and Ishikawa [6] proposed a method involving two stages to minimize the cost of document translation. The translation of queries is done to retrieve results in the target language. The documents of the source language are also translated and arranged based on the translation. Re-ranking helps to improve the precision value of weak query translations.

McCarley [12] and Oard [13] found out that the output of Machine Translation system differs from one language to another. The translations were better when translating for example from French to English or from English to German than when the translation is done from English to French and from German to English. The variation comes due to different morphological analysis.

Hull and Grefenstette [10] compared the monolingual IR system and CLIR systems which translate the queries by using an automatically generated bilingual MRD, or a manually constructed dictionary. MRD proves to be very less efficient as compared to, can lead to a monolingual retrieval. If the correct translations of multi words are found, CLIR system could be as good as a monolingual system.

Pirkola et. al [14] and Bellesteros and Croft [3] bring out the problems of untranslatable words, such as proper names, compound words, and domain specific that are not found in the dictionary used; and inflected words, which could usually be improved by stemming. Xu and Weishedel [17] observed that missing lexicons poses the biggest threat to CLIR system performances.

In the Indian context Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani and Pushpak Bhattacharyya [4] used a query based translation approach and came out with a Hindi to English and Marathi to English CLIR system. D. Thenmozhi, C. Aravindan [15] developed a Tamil-English Cross Lingual Information Retrieval System for Agriculture Society. The farmers of Tamil Nadu retrieved information in English by writing queries in Tamil. Saurabh Varshney, Jyoti Bajpai [16] proposed an algorithm for improving the performance of the English-Hindi CLIR system.

III. APPROACHES IN CLIR

Translation-based CLIR has three major approaches to translation problems: document translation, query translation and interlingual techniques [13]. Document Translation- By applying complete document translation offline, the translations of documents can be obtained. The translation builds up an index for information retrieval and also helps to retrieve the content in native language. Machine Translation has been successfully tried for English, German, Spanish, French etc. but it is not yet available for other pair of languages. In Document translation every document is translated into the language of query and monolingual retrieval is performed. It does not require a passive knowledge of the foreign language from the user. All target languages are translated to the source language. IR process utilizes indexing to enhance the process of search of documents. However indexing cannot be done in case of post translation, so this approach becomes impractical as it uses more time for translation. A long document on the other hand provides more contextual information for translation, which helps to choose the correct options of the terms. However it requires much time to index the collections. Query Translation-There is remarkable improvement in the search techniques to give better results for receiving information for a query. The Search can be refined by providing the intelligent search in the unrestricted domain. Multilingual information search is widely used because information is available online. Query translation helps to find documents in languages different from the language of query. The translation can use the online translation from Google Translate, train a Statistical Machine Translation system using parallel corpora. It uses Machine Readable Dictionaries for translation or use large corpus like Wikipedia [9]. The computational time is less as compared with other methods. Usually a query does not provide enough contexts to automatically find the intended meaning of each term in the query. Translation errors affect the performance significantly. In case of searching a When a multilingual source is searched, query is translated into every language contained in it. Inter-lingual Translation- The inter-lingual representation is performed by combining the methods of document translation and query translation. The Inter-lingual technique is useful if there is no resource for a direct translation but it has lower performance than the direct translation [18]. This approach needs additional storage space for translated documents but provides scalability when same collection of documents is required in multiple languages. These systems use a pre-prepared list of concepts which are language-independent. Then the queries are applied for the same concept in common space. This concept space defines the granularity or precision of possible searching. The main problem of controlled vocabulary systems is that, non-expert users need to be trained. It also require such interfaces which can produce good queries.

The query translation approach involving the translation of input queries is usually efficient and widely used among the three. However, it cannot handle translation ambiguities in the terms because of lack of contextual information. The document translation approach translates documents into the same language as the query. It provides more information on context as documents are longer than queries so the problem of ambiguity can be curbed to some extent. This technique has its limitations as it can only be performed offline because of a large number of documents. Secondly, if the queries are multilingual, the documents need to be translated into more than one language, which increases the burden upon the system. Finally, if there is a change in the material or targets, the whole process of translation has to be repeated.

The inter-lingual technologies attempt to map both queries and documents to a language independent representation. Such instances were created by using multilingual thesauri in the initial stages of CLIR research. But it is no easy task to make a thesaurus, and the automatic mapping of terms in queries and documents to the thesauri is another vast field of research to be addressed separately.

Table: Comparison of three Translation Approaches

Parameter	Query Translation	Document Translation	Inter-lingual Translation
Ambiguity	More	Less	More than both
Additional Storage	Not required	Required	Not required
Translation time	Less	More	More than both
Information Retrieval	Bilingual	Bilingual	Bilingual and Multilingual
Flexibility	High	Less	Less
Working nature	Interface between two language at a time	Interface between two language at a time	Interface between more than two language at a time

IV. CHALLENGES IN CLIR

The major challenges identified in Cross Language Information Retrieval are, Word Inflection and OOV words.

Translation Disambiguation occurs due to the varied meaning of the same word which is termed as homonymy and polysemy [1]. Homonymy refers to a word which has completely different meanings, for example the word "bank" can either mean a river bank or a financial institution. Polysemy refers to a word which has two distinct meanings but are related for example "head" may refer to family head or the human head. There is a need to find out which precise meaning is required in a particular context. It causes ambiguity because a machine cannot interpret meanings like human brain. The translation of a word results in the choice made by the system which may or may not be relevant to the query.

The major issue faced in query translation is of translation ambiguity, and this problem is aggravated because it is difficult to find context in case of short queries. From this perspective, document translation seems to be more capable of producing more precise translation due to richer contexts. The availability of efficient MT systems also makes the document translation approach possible. However, it is not obvious that the current MT systems can take full advantage of the existence of richer contexts in document translation. Several studies have tried to compare the query translation and document translation approaches using the same translation tool. McCarley found that the effectiveness is more dependent on the translation direction between languages than query or document translation: French-to-English translation outperforms English-to-French translation, whether it is used in query translation or document translation. All these experiments show that document translation is not necessarily advantageous to query translation. The main reason behind this observation is that the current MT systems exploit only a limited amount of immediate contextual information, and sentences are usually translated independently. The rich contextual information in documents is largely under exploited and does not significantly impact the quality of the translation. An important point to be identified in document translation is that the language in which the translation is to be done has to be decided beforehand. In a truly multilingual IR environment, one would like to translate each document to all the other languages. This is impracticable because it may require a huge storage space due to the multiplication of document versions. Nevertheless, once a document is pre-translated into the same language as the query, the user can directly read and understand the translated version.

Another problem with query translation is word inflection used in the query. It can be solved by using a Stemmer or a Lemmatizer. In Lemmatization every word is shortened to its base form or lemma; while in Stemming the different forms of a word are reduced to a common shortest form which is called a stem, by removing the endings [11]. In the case of phrases, they cannot be translated word to word but in entirety. The literal word to word translation would bring a change in the meaning [10]. Phrases matched against a manually built multi-word (phrase) dictionary show higher precision than those translated by single word-based dictionaries.

Even in the case of compound words, CLIR faces problems. A compound word is a word formed from two or more words; compound words are not widely available in English, but very much used in other languages. A compound word can be decomposed to two or more words, where each has a meaning are called compositional compounds. The problem arises when non- decomposable compounds whose meaning can't be deduced on the basis of its components, are found in a query.

Using the dictionary-based translation is a traditional approach in cross-lingual IR systems but there are a number of words or terms in the query which appear for the first time and are not included in the dictionary. These words are OOV or Out-of-Vocabulary words.

In many documents, technical terms and proper names are important text elements. Dictionaries only include the most commonly used proper nouns and technical terms used such as major cities and countries. Their translation is crucial for a good cross-language IR system. A common method used to handle untranslatable keywords is to include the untranslated word in the target language query. If a word is not found in the target language, the query may not be able to retrieve the relevant documents [2].So far there have not been any reported research results in cross-language retrieval from document images collections. One problem for cross-language document image retrieval relates to the translation of the output of OCR either for retrieval or content access. A feature of OCR systems is that they make errors in the recognition of individual characters within a word. These errors can sometimes be corrected in post processing, but often they cannot. These recognised "words" are not present in standard dictionaries and thus cannot be translated directly, either by an MT system or by simple dictionary lookup. A method of approximate matching with dictionary entries, perhaps involving steps such as part-of-speech matching and word co-occurrence analysis, might prove effective, but there will remain the possibility of translation errors which result from incorrect word recognition for retrieval or content access.

V. CONCLUSION

CLIR helps in searching documents through different type of languages across the world. It is crucial to the globalized existence. The different approaches have their merits and demerits. Survey indicates that query translation is better than document translation. It is more convenient to translate only the query than the whole documents. Document translation which uses machine translation is computationally expensive and the size of document collection is large. However, it might be more efficient when the computer technology improves. There is no best approach so far yet the research is dedicated to a positive outlook. Hopefully, CLIR would become as popular as search engines in the near future.

VI. REFERENCES

[1]. Abusalah, M., J. Tait,J., and Oakes,M.(2005) "Literature Review of Cross Language Information Retrieval".

[2]. Amelina,N. and Abdullah,M.T."Crosslingual Information Retrieval", Electronic Journal of Computer Science and Information Technology,Vol. 2,No. 1.

[3]. Ballesteros, L. and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In Proceedings of the 20th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Philadelphia, Pennsylvania, United States, July 27 -

31, 1997). N. J. Belkin, A. D. Narasimhalu, P. Willett, and W. Hersh, Eds. SIGIR '97. ACM Press, New York, NY, 84-91.

[4]. Chinnakotla,M., Ranadive,S., Om P. Damani,P. and Bhattacharyya,P.(2004)" Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation"

[5]. Frakes, W. and Baeza-Yates, R. (1992) Information Retrieval: Data Structures and Algorithms. New Jersey: Prentice Hall.

[6]. Fujii, A., Ishikawa, T. (2000). Applying machine translation to two-stage cross-language information retrieval. Proceedings of the 4th Conference of the Association for Machine Translation in the Americas (AMTA-2000), 13-24.

[7]. Grefenstette, G., (1998). The Problem of Cross-Language Information Retrieval, in Cross-Language Information Retrieval, ed. G. Grefenstette Kluwer Academic Publishers, pp. 1–10.

[8]. Hauenschild, C. & Heizmann, S., editors (1997). Machine Translation and Translation Theory Berlin: Mouton de Gruyter.

[9]. Herbert,B, Szarvas, G. and Gurevych,I(2011) "Combining Query Translation Techniques To Improve Cross-Language Information Retrieval".

[10]. Hull, D. A. and Grefenstette, G. (1996). 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 (Zurich, Switzerland, August 18 - 22, 1996). SIGIR '96. ACM Press, New York, NY, 49-57.

[11]. Manning,D., Raghavan,C.P., and Schütze,H. "An Introduction to Information Retrieval", 2009.

[12]. McCarley, J. S. (1999). 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. Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 208-214.

[13]. Oard, D. W. (1998). A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In Proceedings of the Third Conference of the Association For Machine Translation in the Americas on Machine Translation and the information Soup.

Pirkola, A., Hedlund, T., Keskustalo, H., and Jasrvelin, K. (2001). Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. Information Retrieval 4, 3-4, 209-230.

[14]. Thenmozhi,D. and C. Aravindan ,C.(2009)"Tamil-English Cross Lingual Information Retrieval System for Agriculture Society".

[15]. Varshney,S. and Bajpai,J.(2007) "Improving performance of English-Hindi Cross Language Information Retrieval using Transliteration of query terms"

[16]. Xu, J., Weischedel, R., and Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval, 105-110.

[17]. Zhou ,D., Truran ,M., Brailsford,T., Vincent Wade,V. and Helen Ashman,H.(2012)" Translation Techniques in Cross-Language Information Retrieval".