Cross-Language Information Retrieval on Indian Languages: A Review

Nitin Verma*
Suket Arora**
Preeti Verma***

Abstract

Cross Language Information Retrieval on Indian Languages (CLIROIL) can be used to improve the ability of users to search and retrieve documents in different languages. The aim of CLIR is to provide the benefit to the user in finding and assessing information without being limited by language barriers. We can use Simple measures to get high - accuracy in cross-language retrieval in which translation is one of them. Translation is one of the technique that makes use of software that translates text from one language to another language. Different type of translation techniques (dictionary based translation, machine translation, transitive translation, dual translation) can be used to achieve Cross Language Information Retrieval. IR deals with presentation, storage, space, retrieval, and access of a multiple document collection. This paper describes the work done in CLIR and translation techniques for CLIR. This paper translates the work done.

Keywords: CLIROIL, Translation, Dictionary-based, Machine translation, Transitive translation.

I. Introduction

Cross Language Information Retrieval On Hindi Language allows the users to read and search pages in the language different from the other language of being searched. Cross language information retrieval is a kind of information retrieval in which the language of the query is different from the language of the documents retrieved as in a search result. In Cross Language Information Retrieval system a user is not limited to his own native language, different set of languages are there, so the user can make his query in his native language but the system returns set of documents in another different languages. Different foreign languages have been used like English, French, Spanish, Chinese. But Indian languages always have Cross Language Information Retrieval On Hindi Language allows the users to read and search pages in the language different from the other language of being searched. Cross language information retrieval is a kind of information retrieval in which the language of the query is different from the language of the documents retrieved as in a search result. In Cross Language Information Retrieval system a user is not limited to his own native language, different set of languages are there, so the user can make his query in his native language but the system returns set of documents in another different languages. Different foreign languages have been used like English, French, Spanish, Chinese. But Indian languages always have system simplifies the search process for multiple users and enables those who know only one language to provide queries in their language and then get help from translators for using other languages documents. CLIR system simplifies the search process for multiple users and enables those who know only one language to provide queries in their language and then get help from translator for using other languages documents. CLIR. System simplifies the search process for multiple users and enables those who know only one language to provide queries in their language and then get help

Nitin Verma*
Assistant Professor, Computer Science Dept., Hindu College, Amritsar

Suket Arora**
Assistant Professor, Dept. of Computer Applications, Amritsar College of Engineering & Technology, Amritsar

Preeti Verma***
Assistant Professor, Dept. of Computer Applications, Amritsar College of Engineering & Technology, Amritsar

from translators for using other languages documents. Due to the "standardization" of terms, stemming sometimes contributes in increasing the retrieval effectiveness. This is, however, not always the case. Current search engines usually do not use aggressive stemming, while in the area of research, stemming is still generally used as a standard pre-processing.

II. Translation

A full document translation can also be applied offline to create translation of an entire document. The translations provide the basis for constructing an index for information retrieval and also offer the user the possibility to access the content in his native language. Multiple information search becomes important due to large amount of online information available in different languages. We can also use an online translation through sources like i.e. Google, Wikipedia which confirms the accuracy of the search. Usually machine translation system supports the translation. Searching strategies are continuously improving their techniques to provide more relevant, accurate and proper information for a given query. A common problem with translation is word accuracy. This problem can be solved by using different techniques .Various techniques are used to reduce the grammatical mistakes. The Search can also be filtered by providing the unrestricted domains. Machine Translation is not always available as a realistic option for every pair of languages. Widely translation system supports the translation between language pairs which involve the languages likely as English, German or Spanish, and Chinese. In translating the document, firstly we select a single query language and then translate every single document into that language then single retrieval is carried out. This technique provides more context but current systems don't damage the context widely. But one must have to determine in which language each document should be translated; translated documents in all the languages should be stored.

III. Translation Techniques

Translation techniques in CLIR are categorized into two types:

- Direct translation
- Indirect translation

A. Direct Translation

The direct is of three types. Now we will explain them:

- Corpus Based Translation
- Dictionary Based Translation
- Machine Based Translation

1) Corpus Based Translation

Parallel corpora are commonly used in cross-language information retrieval to translate queries. The basic technique involves a side-by-side analysis of the corpus producing a set of translation probabilities for each term in a given query[1]. Large collections of parallel texts are referred to as parallel corpora. Parallel corpora can be acquired from a variety of sources.

2) Dictionary Based Translation

A dictionary-based approach for the translation is very easy but it is having two limitations such as ambiguity and lack of coverage[1].

3) Machine Translation

Machine Translation is not only performs the substitution of words from one language to other; but it also involves finding phrases and its counterparts in target language to produce good quality translation.

B. Indirect Translation

Indirect translation relies upon the use of an intermediary which is placed between the source query and the target document collection. In the case of transitive translation, the query will be translated into an intermediate to enable comparison with the target document collection. The Indirect translation is two types:

- Transitive translation
- Dual translation

1) Transitive Translation

Transitive translation relies upon the use of a pivot language which acts as an intermediary between the source query and the target document collection[1].

2) Dual Translation

Dual translation systems attempt to solve the query document mismatch problem by translating the query representation and the document representations into some "third space" prior to comparison. This "third space" can be another human language, an abstract

language or a conceptual inter-lingual. This general category also includes translation techniques that induce a semantic correspondence between the query and the documents in a cross-language dual space defined by the documents.

IV. Approaches of clir

There are different approaches for CLIR. Following are approaches:

A. Query Translation

Multilingual information search becomes important due to increasing the amount of online information available in non-English languages and multiple language document collections. This can be achieved by Query translation. Query translation using CLIR became the widely used technique to access documents of the different languages from the language of query. For translating the query, we can use an online translation i.e. Google Translate, train a Statistical Machine Translation system using parallel corpora, employ Machine Readable Dictionaries to translate query terms or use of large scale multilingual information sources like Wikipedia . Google Translate query translation approach. Translation can be applied to the query terms online. Online query translation can be achieved by using one of the Google Translate API which will convert the query into the other languages. Online query translation will help the user to translate his query in the other languages. Online query translation will help the user to translate his query in the other languages [3].

B. Interlingual 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. The Inter-lingual technique is useful if there is no resource for a direct translation but it has lower performance than the direct translation [4].

C. Document Translation

In Document translation we select a single query language and then translate every document into that language then perform monolingual retrieval. Typically machine translation systems supports the translation between language pairs which involve languages, such as English, German or Spanish, and English.

D. Some Advance Approaches

1) Universal words

They confirm the vocabulary of the language. To be able to express any concept occurring in a natural language, the UNL proposes the use of English words modified by a series of semantic restrictions that eliminate the innate ambiguity of the vocabulary in natural languages. If there isn't any English word suitable to express the concept, the UNL allows the use of words from other languages. In this way, the language gets an expressive richness from the natural languages but without their ambiguity.

2) Relations

These are a group of 41 relations that define the semantic relations among concepts. They include argumentative (agent, object, goal), circumstantial (purpose, time, place), logic (conjunction, and disjunction) relations, etc.

V. Knowledge Representation

By knowledge bases in our context we understand the set of concepts belonging to a specific domain and the relations between these concepts that also belong to this domain. But when we turn to ontologies, the richness of a domain becomes relegated to a mere enumeration of concepts and a taxonomic organization of them. That is, there is danger of identifying ontologies as mere theasauri.[8]

VI. Challenges In CLIR

- 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 non-translated word in the target language query. A phrase cannot be translated by translating each of the word in the phrases.
- Named entities extraction and translation are vital in the field of natural language processing for research on machine translation, cross language IR, bilingual lexicon construction, and so on. There are three types of Named entities; entity names such as organizations, persons and

locations, temporal expressions such as dates and times, and number expressions such as monetary values and percentages.

- 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. This is called the Out-of-Vocabulary. This is to be expected even in the best of dictionaries. Translation Disambiguation, which is rooted from homonymy and polysemy[6]. Homonymy refers to a word which has at least two entirely different meanings, for example the word "left" can either mean opposite of right or the past tense of leave. Input queries by user usually short and even the query expansion cannot help to recover the missing words because of the lacking information.[7]
- A common problem with query translation is word inflection used in the query. This problem can be solved by stemming and lemmatization. Lemmatization is where every word is simplified to its uninflected form or lemma; while stemming is where different grammatical forms of a word are reduced to a common shortest form which is called a stem, by removing the ending in word. For example, the stemming rules for word "see" might return just "s" by stemming and "see" or "saw" by lemmatization[4].

VII. Applications of CLIR

- This CLIR System can be helpful for immigration department. For eg. Immegration department interact with thousands of the Indian native Language speakers which are not able to understand English Languages .
- This System can be used for multilingual population regions so that the peoples having different native languages retrieve documents in their native languages.
- This system can also be used for intelligence departments.
- The CLIR will be beneficial for students for their research work regarding historical places.

VIII. Conclusion

CLIROIL provides us a new technique for searching documents through different kinds of languages across the whole world .By using the different type of translation techniques CLIROIL make it possible to provide the better search results in the other language to the language which is queried. So it will be beneficial for wide population regions. Survey proves that query translation is much better than document translation. It is more convenient way to translate the query than the whole documents. Document translation which uses machine translation is computationally quite expensive and the size of document collection is large. However, it might be practical in the future when the computer technology would be much improved.

References

1. Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, Helen Ashman," Translation Techniques in Cross-Language Information Retrieval.
2. J. Cardeñosa, C Gallardo, Adriana Toni," Multilingual Cross Language Information Retrieval A new approach".
3. UNL Center. UNL specifications v 2005. http://www.undl.org/unlsys/unl/unl2005-e2006/
4. D. Manning, C., P. Raghavan, and H. Schütze, "*An Introduction toInformation Retrieval",* 2009.
5. Nurul Amelina, Nasharuddin, Muhamad Taufik Abdullah, "Crosslingual Information Retrieval",Electronic Journal of Computer Science and Information Technology,Vol. 2,No. 1.
6. Abusalah, M., J. Tait, M. Oakes, "Literature Review of Cross Language Information Retrieval",2005
7. Nurul Amelina, Nasharuddin, Muhamad Taufik Abdullah,"Crosslingual Information Retrieval",Electronic Journal of Computer Science and Information Technology,Vol. 2,No. 1,
8. Bateman, J.A; Henschel, R. and Rinaldi, F. "The Generalized Upper Model 2.0." 1995. http:// http:// www.fb10.unibrem en.de/anglistik/langpro/webspace/jb/gum/index.htm