ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

A Survey on Cross Language Information Retrieval

Monika Sharma[1], Sudha Morwal[2]

Department of Computer Science, Banasthali University Jaipur, India[1,2]

Abstract: Search for the information is not only limited to the native languages of the user, but nowadays it is more extended to other languages. Cross language information retrieval (CLIR), whose goal is to find relevant information written in a language different from the language of query. CLIR can be used to enhance the ability of users to search and retrieve documents in many languages. Different type of translation techniques can be used to achieve Cross Language Information Retrieval. This paper describes the work done in CLIR and translation techniques for CLIR

Keywords: Query translation, Document translation, CLIR, Machine translation, Dictionary

I. INTRODUCTION

CLIR is the task of issuing a query in one language and retrieving the set of relevant documents in other languages. The aim of CLIR is to provide the benefit to the user in finding and assessing information without being limited by language barriers. CLIR has become more important in recent years. CLIR enables the users to retrieve documents of the other languages than their original query language. Potential users for CLIR are users who find it difficult to formulate a query in their non-native language and users who are multilingual and want to save time by entering a query in one language instead of entering the query in all languages in which they want , Cardeñosa.J.et al, [1].

Cross-Language Information Retrieval allows the users to search and read pages in the language different from the language of the search terms Swapna.N.et al, [2]. Cross-language information retrieval is a type of information retrieval in which the language of the query is different from the language of the documents retrieved as a search result. In CLIR system a user is not limited to his own native language, so the user can make his query in his native language but the system returns set of documents in another languages.

CLIR involves the retrieval of documents in a language other than the query language. Since the language of query and the documents needs to be translated in CLIR. CLIR system simplifies the search process for multilingual 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, Alizadeh.H.et al, [3].Cross language information retrieval make it possible to retrieve the multilingual documents.

II. WORK DONE IN CLIR

1. Nandkishor Vasnik 2012 [4], in his research use the NLP techniques approaches for improving the quality of the search on Internet. In which query extensions and improving the quality of information retrieved using NLP-based systems. The main goal of system, TALASH: A Hindi Search Engine is to improve the result provided by Google search engine through the extension of user Hindi input. Result of search engine is depend on database present and how structured is it. This Research aims to provide linguistic mechanisms that transform and extend the user query by integrating Hindi Word Net semantic database, and user context

2.Saurabh Varshney,Jyoti Bajpai [5] in their research proposed an algorithm for improving the performance of the English-Hindi CLIR system. He try to use all possible combination of Hindi translated query using transliteration of English query terms and choosing the best query among them for retrieval of documents.

3.B. Herbert, G. Szarvas, I. Gurevych [6] in their research presents the combination of query translation approaches for cross-language information retrieval (CLIR). They translate queries with Google Translate and extend them with new translations obtained by mapping noun phrases in the query to concepts in the target language using Wikipedia.

4. D. Thenmozhi, C. Aravindan [7] presents a Tamil-English Cross Lingual Information Retrieval System for Agriculture Society. In their research, they developed a CLIR system in Agriculture domain for the Farmers of Tamil Nadu which helps them to specify their information need in Tamil and to retrieve the documents in English. Local word reordering is performed according to Subject-Verb-Object pattern in order to preserve the relative dependency across the words. Word sense disambiguation is performed that identifies the correct sense of an ambiguous word that is being used in a query.

5. Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani and Pushpak Bhattacharyya[8] presents a Hindi to English and Marathi to English CLIR system. They use a query based translation approach using dictionaries. Query words that are not found in the dictionary are translated using a simple rule based translation technique. The resultant translation is then compared with the unique words of the corpus to return the ‘k’ words most similar to the translated word.

DOI 10.17148/IJARCCE.2015.4287

6. Debasis Mandal, Sandipan Dandapat, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar[9] presents a Bengali and Hindi to English CLIR system. The cross-language task includes the retrieval of English documents in response to queries in two most widely spoken Indian languages, Hindi and Bengali. They use automatic Query Generation and Machine Translation approach .

7. Mallamma V Reddy, Dr. M. Hanumanthappa [10] presents Kannada English and Telugu English CLIR systems as part of Ad-Hoc Bilingual task. Kannada and Telugu Native Languages to English. They use a query translation based approach using bi-lingual dictionaries. When a query word is not found in the dictionary then the words are translated using a rule based approach which utilizes the corpus to return the 'k' closest English transliterations of the Kannada/Telugu word. The resulting multiple translation choices for each query word are disambiguated using an iterative page-rank style algorithm which is based on term-term co-occurrence statistics and then produces the final translated query.

III. APPROACHES IN CLIR

A. DOCUMENT TRANSLATION

Full document translation can be applied offline to produce translations 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. Machine Translation is not always available as a realistic option for every pair of languages. Typically machine translation systems supports the translation between language pairs which involve languages, such as English, German or Spanish, and English [1].

In Document translation we select a single query language and then translate every document into that language then perform monolingual retrieval. This approach provides more context but current MT systems don't exploit the context much.But one must have to determine in which language each document should be translated; translated documents in all the languages should be stored.

B. .QUERY TRANSLATION APPROACH FOR CLIR

Search strategies are continuously improving their techniques to provide more relevant, accurate information for a given query. The Search can be refined by providing the intelligent search in the unrestricted domain. 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 . CLIR using the Google Translate achieve the accuracy of the search results up to 90 %, Herbert.B.et al [11]. Query Translation Mapping the query representation into the document representation .

We can use 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 [1]. This approach is more flexible because it involves more interaction with the user .Users can choose languages of their interest and also correct translated query. But there is translation ambiguity due to lack of context.

C. INTERLINGUAL TRANSLATION

The Interlingual representation is the combination of the document and query translation techniques. The Interlingual technique is useful if there is no resource for a direct translation but it has lower performance than the direct translation [12].

IV. TRANSLATION TECHNIQUES

Translation techniques in CLIR are categorized into direct translation and indirect translation.

A. DIRECT TRANSLATION

Direct translation systems uses the bilingual dictionaries, parallel corpora and machine translation algorithms to translate the source text. We discuss each of them below:

1.) Dictionary-Based Translation

Machine-readable bilingual dictionaries have become increasingly available and are often used in the translation modules of CLIR engines. A dictionary-based approach for the translation is very easy but it is having two limitations such as ambiguity and lack of coverage [12].

2.) Machine Translation

Machine translation (MT) is the automatic translation of the text from one natural language to another. MT systems have become popular in CLIR due to the wide availability of MT systems and the linguistic resources required to train them. Machine translation, is a technique that makes use of software that translates text from one language to another language. 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 translations [12].

3.) Corpus-based Translation

A parallel text is a document written in one language together with its translation in another language. Large collections of parallel texts are referred to as parallel corpora. Parallel corpora can be acquired from a variety of sources. International organizations such as the United Nations and the European parliament17 publish a huge volume of parallel documentation every year in a wide variety of languages. 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 [12].

DOI 10.17148/IJARCCE.2015.4287

B. INDIRECT TRANSLATION

Indirect translation is a common solution when there is a absence of resources supporting direct 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. In the case of dual translation systems, both the query and the document representations are translated into the intermediate language [12].

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 [12].

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 interlingua. 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 [12].

V. CHALLENGES IN CLIR

Each of the approaches has challenges to Cross Language Information Retrieval.

1. Translation Disambiguation, which is rooted from homonymy and polysemy [13]. 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. Polysemy refers to a word which can take on two distinct but related meanings such as the "head" of the family or the human's "head". So there becomes a problem when finding the most appropriate translation from various choices in the dictionary. Very frequently, translation of a word results in such a choice having to be made.

2 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 [14].

3 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. Input queries by user usually short and even the query expansion cannot help to recover the missing words because of the lacking information [15].

4 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 this word does not exist in the target language, the query will be less likely to retrieve the relevant documents. Translating phrases is also becoming one of the problems in cross-lingual IR. A phrase cannot be translated by translating each of the word in the phrases [15].

5 Named entities are essential components of texts, especially news texts [15]. 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.

VI. APPLICATIONS

1.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 .

2.This System can be used for multilingual population regions so that the peoples having different native languages retrieve documents in their native languages.

3.This system can also be used for intelligence departments.

4.The CLIR will be beneficial for students for their research work regarding historical places.

VII. CONCLUSION

CLIR provides new technique for searching documents through different type of languages across the world .By using the different type of translation techniques CLIR make it possible to provide the search result in the other language to the language of query. So it will be beneficial for multilingual population regions. 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 practical in the future when the computer technology improves

REFERENCES

[1] J. Cardeñosa, C Gallardo, Adriana Toni," Multilingual Cross Language Information Retrieval A new approach".

[2] N.Swapna, N.Hareen Kumar, B.Padmaja Rani,"Information Retrieval in Indian Languages: A case study on Cross-Lingual and Multi-Lingual", International Journal of Research in Computer and Communication technology ,ISSN 2278-5841,Vol 1,Issue,September 2012.

DOI 10.17148/IJARCCE.2015.4287

[3] H. Alizadeh, R. Fattahi "Applying Natural Language Processing Techniques for Effective Persian- English Cross-Language Information Retrieval", International Journal of Information Science and Management.

[4] N. Vasnik, S. Sahu, D. Roy,"TALASH: A Semantic and context based optimized Hindi search engine", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[5] S. Varshney, J. Bajpai, "Improving performance of English-Hindi Cross Language Information Retrieval using Transliteration of query terms"

[6] B. Herbert, G. Szarvas, I. Gurevych "Combining Query Translation Techniques to Improve Cross-Language Information Retrieval"

[7] D. Thenmozhi, C. Aravindan "Tamil-English Cross Lingual Information Retrieval System for Agriculture Society".

[8] Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani and Pushpak Bhattacharyya," Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation".

[9] D. Mandal, S. Dandapat, M. Gupta, P. Banerjee, S.Sarkar, "Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources", At the 8th Workshop of the Cross-Language Evaluation Forum, Budapest, Hungary, 19-21 September 2007.

[10] Mallamma V Reddy, Dr. M. Hanumanthappa," Kannada and Telugu Native Languages to English Cross Language Information Retrieval ",International Journal of Computer Science and Information Technologies, Vol. 2.

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

[12] Dong Zhou,Mark Truran,Tim Brailsford, Vincent Wade,Helen Ashman," Translation Techniques in Cross-Language Information Retrieval".

[13] Abusalah, M., J. Tait, M. Oakes, "Literature Review of Cross Language Information Retrieval",2005.

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

[15] Nurul Amelina,Nasharuddin,Muhamad Taufik Abdullah,"Cross-lingual InformationRetrieval",Electronic Journal of Computer Science and Information Technology,Vol. 2,No. 1,

DOI 10.17148/IJARCCE.2015.4287