Analysis of Cross Language Information Retrieval methods

Ketan Rajshekhar Shahapure

CMSC 676
Information Retrieval

Term paper

ABSTRACT

Information Retrieval systems should be capable of searching for information in multiple languages. This research area called Cross Language Information Retrieval (CLIR) is an intersection of Machine Translation and Information Retrieval. There are 3 problems of CLIR – knowing how a term can be converted to another language, which of the possible translations should be retained and how to properly weigh the importance of translation alternatives.

There are two approaches for finding translations between different languages. First one is using a bilingual dictionary. Bilingual dictionaries already exist in a machine-readable format for many different languages. But these dictionaries also have many problems associated with them like missing word forms where a word like ‘graduate’ is present, but its past tense ‘graduated’ might be missing. The second approach is using parallel corpus in which same text is written in different languages. If the parallel corpus is large enough then simple statistical techniques can be used to produce bilingual equivalents of terms.

The focus would be on covering the recent trends in Cross Language Information Retrieval research. The trends analyzed are Keizai CLTR system for English to Japanese or Korean translation, English – Hindi CLIR system, Cross Language Information Retrieval and Delivery using community mobile networks for English to Tamil translation and other south Indian languages and using ontologies for English to Persian translation.

1. INTRODUCTION

CLIR is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query. Information Retrieval systems should be capable of searching for information in multiple languages. Cross Language Information Retrieval (CLIR) is an intersection of Machine Translation and Information Retrieval.

Cross Language Information Retrieval is created to recover content reports in a language not the same as the language used to determine the data required. Search engines are assuming indispensable job to web clients in finding the destinations they need to visit and content they need to browse. Web crawlers give the capacity of making search queries in different nearby dialects and even they decipher the corpus from English into different nearby dialects and the other way around.

2. MOTIVATION FOR CLIR SYSTEMS

The motivation for developing Cross Language Information Retrieval systems is the need to acquire information even if it's not available in the user's native language. CLIR may bridge the gap between the desire to obtain information and unavailability or under-availability of such information in their native language. It helps retrieve information from a multilingual collection using a query in a single language. CLIR is also used to locate documents in a multilingual collection of scanned pages.

Cross Language Information Retrieval research is important for global information exchange and sharing of knowledge. CLIR finds its uses in the following areas:

- National Security
- Foreign Patent information access
- Medical information access for patients
- Sentiment analysis
- Information Extraction

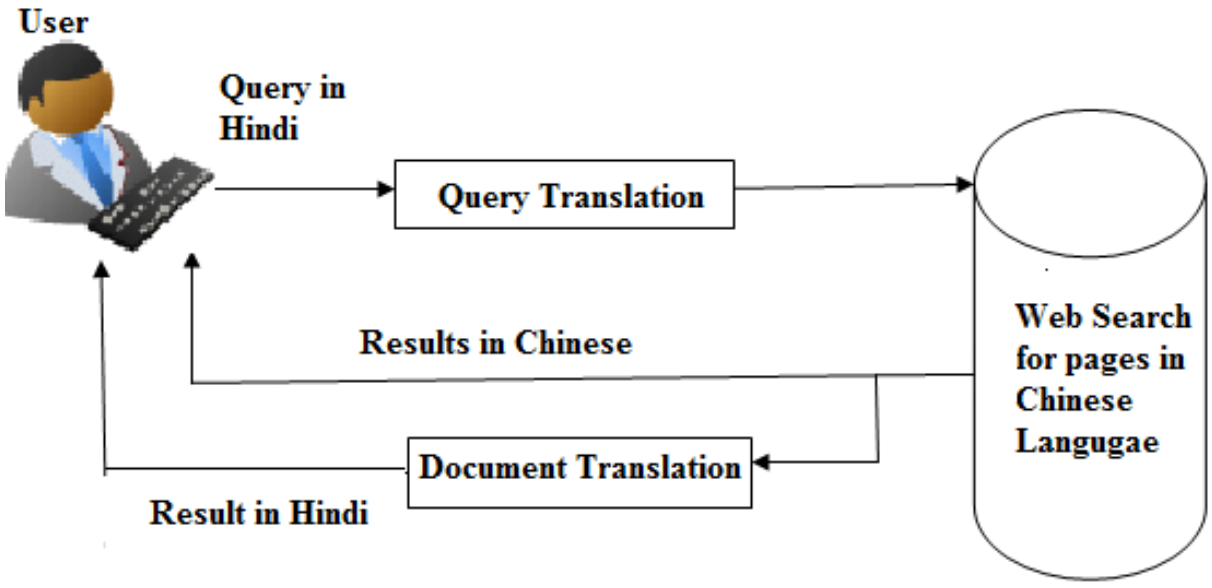

Figure 1. Example of CLIR system

Dwivedi, Sanjay & Chandra, Ganesh. (2016). A Survey on Cross Language Information Retrieval. International Journal on Cybernetics & Informatics. 5. 127-142.

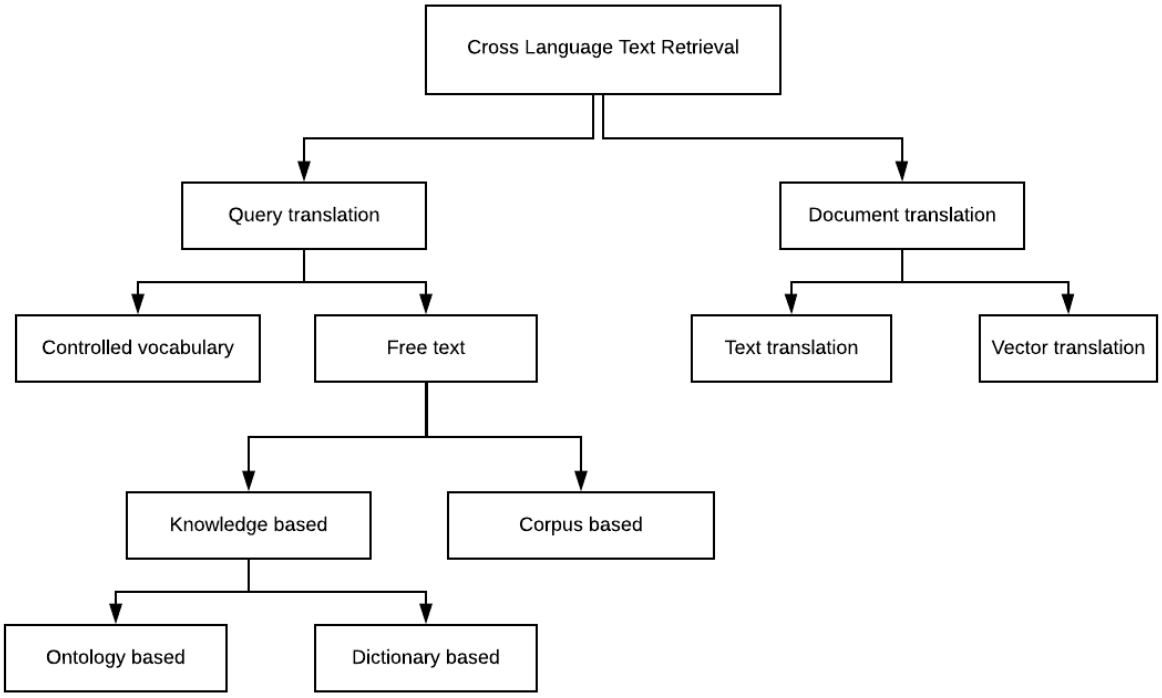

Figure 2. Types of Cross Language Text Retrieval techniques

3. ISSUES AND DESIGN DECISIONS FOR CLIR SYSTEMS

There are 3 main issues encountered while developing a Cross Language Information Retrieval system. They are as follows:

- How to convert a term to another language?
- Which of the possible translations should be retained?
- How to properly weigh the importance of translation alternatives?

Important design decisions also need to be made while developing such Cross Language Information Retrieval systems. They are as follows:

- What to index?
    - Free text or controlled vocabulary
- What to translate?
    - Queries or documents
- Where to get translation knowledge?
    - Dictionary, ontology, training corpus

4. RECENT TRENDS IN CLIR RESEARCH

A number of conferences, tracks, workshops are being conducted to carry out research in the CLIR area. Among those The Cross-Language Evaluation Forum (CLEF) is an activity of the DELOS Network of Excellence for Digital Libraries. The goal of CLEF is to provide an evaluation infrastructure and benchmarking facilities for the testing and tuning of monolingual and cross-language information retrieval systems operating on European languages. This paper analyses the following CLIR systems

- Keizai CLTR system
- English – Hindi CLIR system
- Cross Lingual Information Retrieval and Delivery using community mobile networks
- Ontologies

4.1 Keizai CLTR system

The Keizai CLTR system is used for translation from English to Japanese / Korean and vice-versa. This system uses the query translation approach. The user of this system inputs query in English language and the system searches Japanese and Korean web data. It then displays a list of documents and shows English summaries on top ranking documents. The system is not completely autonomous and user intervention is needed and user needs to accurately judge which foreign language documents are relevant to their query. The system also provides extended English definitions of query terms alongside Japanese or Korean translations.

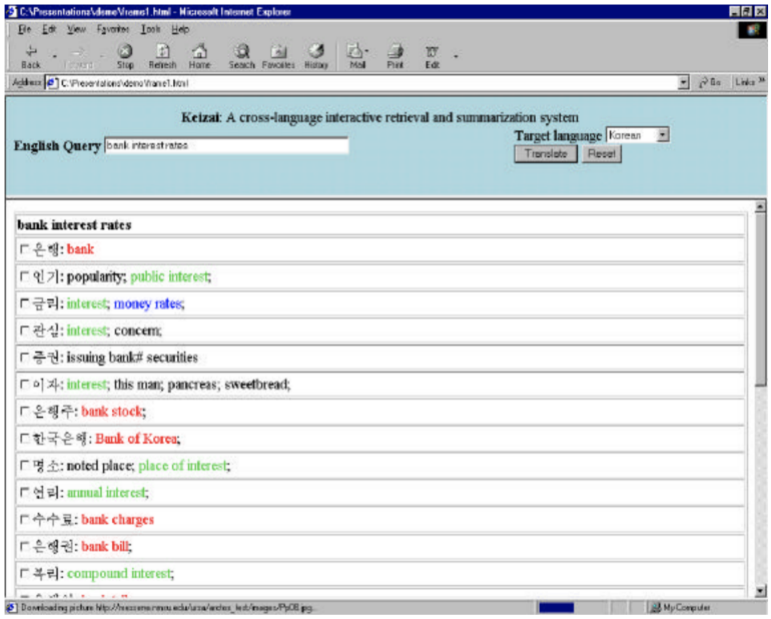

Figure 3. Keizai Query term selection

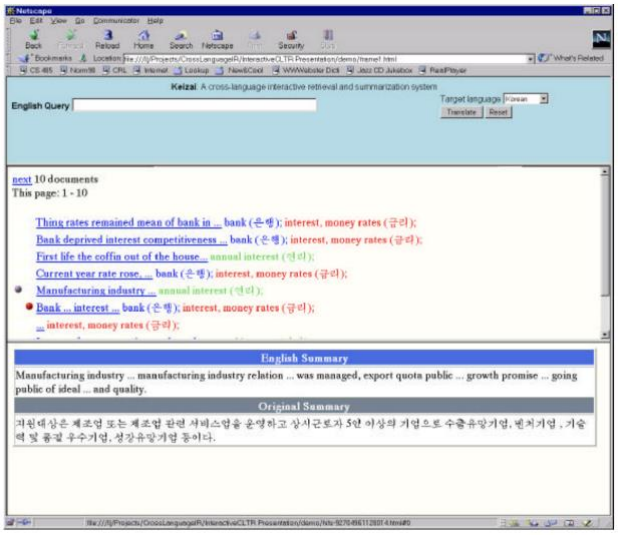

Figure 4. Keizai Display of result documents

4.2 English – Hindi CLIR system

This CLIR system is developed using Managing Gigabytes (MG) retrieval system as the base IR system. It converts the query from user in English to Hindi language. Publicly available online bilingual dictionary ‘Shabdanjali’ was used for query translation while developing this system. The quality of translation in this system depends on the quality of bilingual dictionary used.

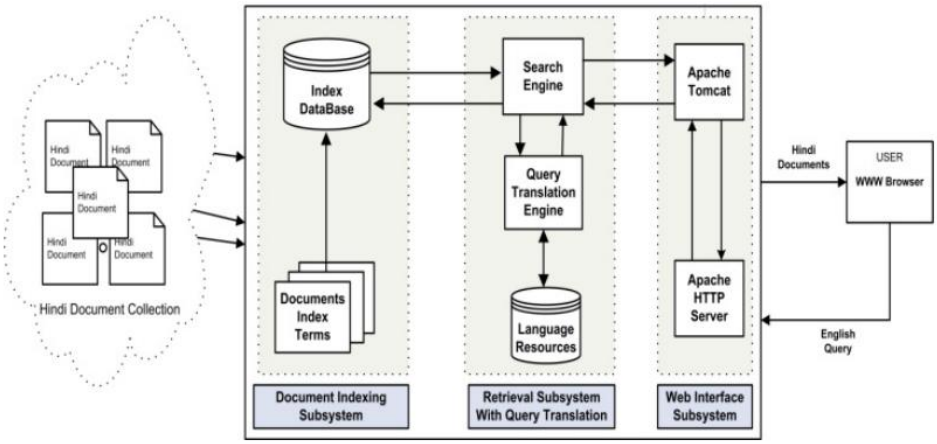

Figure 5. English - Hindi CLIR system

4.3 CLIR and delivery using community mobile networks

This system Focuses on querying the web in languages other than English, namely south Indian languages including Tamil. It searches the appropriate content and summarizes using a content-specification meta language developed as part of the research. The system retrieves relevant documents, translates, summarizes and presents the information to user in Tamil language. A few endeavors have been attempted as for utilizing Tamil for social media on the Internet. This examination expands on the outcomes in the regions of syntactic parsing of Tamil and Tamil web crawler in building up the CLIR framework for Tamil and other south Indian dialects that interface with web search tools like Google, Yahoo, Bing and DuckDuckGo.

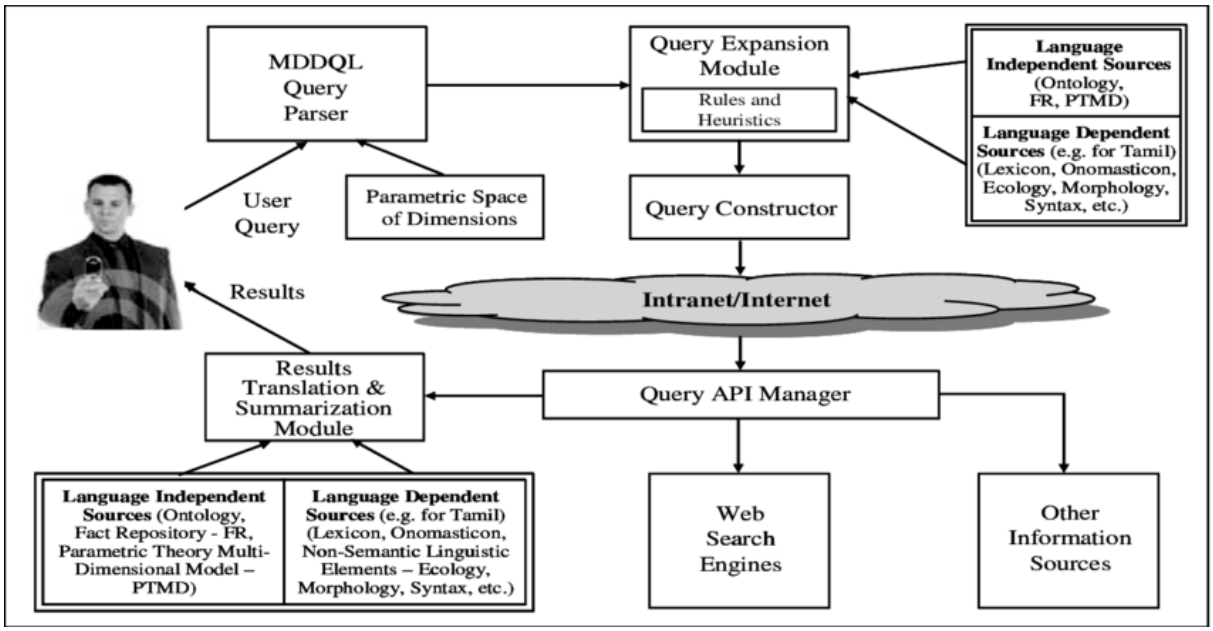

Figure 6. Architecture of CLIR using community mobile networks

4.4 ONTOLOGIES

Ontology is a formal, explicit specification of a shared conceptualization. This system retrieves English documents relevant to Persian queries using Bilingual ontology to annotate the documents and queries. A bilingual ontology consists of ontology and a bilingual dictionary. Ontology is used to expand the query with related terms in pre and post translation expansion and the combined approach significantly improves cross-lingual performance. The researchers of this system analyzed query translation in cross lingual IR based on feature vectors and usage of context information. It was observed that using information external to the query, such as the ontologies, the effect of disambiguation can be reduced.

5. FUTURE SCOPE OF CLIR SYSTEMS

5.1 AVAILABILITY FOR ALL LANGUAGES

As of now a large portion of the cross-lingual research includes only the top generally utilized languages on the planet for instance English, Chinese, French, Spanish and Hindi. Additionally, investigation has being done on dialects that have being impacted most by the finance and business of a nation. That means CLIR systems have been developed only for a few of these popular languages and other less popular languages are left out. This means that the people who only know the excluded languages are not able to access the information in other languages. There is need for research and development of CLIR systems for these languages.

5.2 MULTI-LINGUAL IR

When there are numerous languages accessible to be utilized for looking through data, it will raise the issue of multilingual IR. This kind of IR isn't confined to just two dialects however can

likewise incorporate the same number of dialects as the framework can. This will in the long run expand the list items that are recovered which are pertinent to the questions from the clients and upgrade the framework's viability. In any case, then again, clients will need to invest energy and perhaps cash to physically interpret all the related reports to their primary language.

6. CONCLUSION

Cross-lingual IR gives new ideal models in looking through various languages over the world and it can be the benchmark for looking not just among two languages, but it can be used for searching multiple languages. This paper clarifies a depiction on cross lingual IR, its difficulties and current strategies also, procedures to overcome issues for effective and resourceful searching.

REFERENCES

[1] Ogden, William & Cowie, James & Davis, Mark & Ludovik, Eugene & Nirenburg, Sergei & Sharples, Nigel. (2000). Keizai: An Interactive Cross-Language Text Retrieval System.

[2] Raghunathan, Shriram & Sugumaran, Vijayan & Kapetanios, Epaminondas. (2007). Cross-Lingual Information Retrieval and Delivery Using Community Mobile Networks. 320 - 325. 10.1109/ICDIM.2007.369217.

[3] A. Seetha, S. Das and M. Kumar, "Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method," 10th International Conference on Information Technology (ICIT 2007), Orissa, 2007, pp. 56-61.

[4] V. Pemawat, A. Saund and A. Agrawal, "Hindi - English based cross language Information Retrieval system for Allahabad Museum," 2010 International Conference on Signal and Image Processing, Chennai, 2010, pp. 153-157.

[5] B. A. Kumar, "Profound Survey on Cross Language Information Retrieval Methods (CLIR)," 2012 Second International Conference on Advanced Computing & Communication Technologies, Rohtak, Haryana, 2012, pp. 64-68.

[6] Jian-Yun Nie, "Cross-Language Information Retrieval," in Cross-Language Information Retrieval , Morgan & Claypool, 2010

[7] P. Liu, Z. Zheng and Q. Su, "Cross-Language Information Retrieval Based on Multiple Information," 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, 2018, pp. 623-626.