A Review on the Cross and Multilingual Information Retrieval

PothulaSujatha and P. Dhavachelvan

Department of Computer Science, Pondicherry University, Puducherry-605014, India
{spothula, dhavachelvan}@gmail.com

Abstract:

In this paper we explore some of the most important areas of information retrieval. In particular, Cross-lingual Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR). CLIR deals with asking questions in one language and retrieving documents in different language. MLIR deals with asking questions in one or more languages and retrieving documents in one or more different languages. With an increasingly globalized economy, the ability to find information in other languages is becoming a necessity. We also presented the evaluation initiatives of information retrieval domain. Finally we have presented the overall review of the research works in Indian and Foreign languages.

Keywords:

Cross-Lingual Information Retrieval, Dictionary Based Translation, Machine Translation, Ontology.

1. Introduction

Information Retrieval [IR] is the act of storing, searching, and retrieving information that match a user's request [3]. With the start of the Internet, information retrieval became increasingly relevant and researched. Now, most people use some type of modern information retrieval system on a daily basis, whether it is Google or some specially created system for libraries. This deals with asking question in one language and retrieving documents in one or more different languages. The variants of the IR are BLIR, CLIR and MLIR. In this paper we will only concentrate on CLIR and MLIR. CLIR deals with asking questions in one language and retrieving documents in different language. MLIR deals with asking questions in one or more languages and retrieving documents in one or more different languages. With an increasingly globalized economy, the ability to find information in other languages is becoming a necessity.

The rest of the paper is organized as following. Section 2 introduces the most important IR methods, various works of CLIR and its future is presented in Section 3, and in Section 4 MLIR system and review its previous research works are discussed, at last, we conclude in Section 5.

2. IR Methods

In classical IR search engines, both the query and the retrieved documents are in the same language. The classical IR regards the documents in foreign language as the unwanted "noise" [1]. These needs to introduce new area of IR which takes into account all the documents received regardless of the languages being used. This is where the bilingual, cross-lingual and multi-lingual IR plays a part. But to perform these variants of IR, a variety of translation methods are required. These are described in the following sub-sections.

DOI : 10.5121/ijwest.2011.2409

Translation can be done to the query, the document or both when any retrieval system involved with many languages. Query translation involves translating the query to the target language. Document translation translates the document into the source language (i.e. the language used for the query). There are various methods to translate query, document or both. There are three primary tools for translations are dictionaries, machine translation systems and parallel corpora. Query translation, typically, uses either dictionary based or corpus based translation. Document translation, for the most part, only uses machine translation.

2.1 Dictionary Based

A bilingual dictionary is a list of words in the source language and their translation(s) in the target language. Optionally, these dictionaries have translation probabilities assigned that allow for disambiguation and weighting. There are plenty of bilingual dictionaries are available in the literature both in Indian and Foreign languages.

2.2 Machine Translation

The Machine translation method simply uses a machine translation system to translate either the document or query. The main drawback of this method is computational expensive. In situations where there is a large collection of documents or when searching for documents on the web, machine translation is impractical.

2.3 Parallel Corpora

When compared to dictionary based corpus based translation typically gives much better performance, as [5] found. However, the creation of parallel corpora is complicated and quite expensive. It can be extremely difficult to find parallel corpora for certain languages or that are large enough to be of use.

The main problems with both corpus based and dictionary based translation are coverage and quality. Poor quality corpora and dictionaries can greatly decrease the performance of a system [5]. Coverage relates to out of vocabulary words, or words that are not present in the dictionary or corpus. These words will have no translation, while in some languages that are related this is no problem in other language pairs such as Chinese and English this is a big problem [6]. Because of this there has been considerable research done on automatically or semi-automatically acquiring parallel corpora or bilingual lexicons.

The same methods are used for CLIR and MLIR. These two systems may use translation of all documents into a common language, either automatic translation of the queries, or combination of both query and document translations.

3. CLIR

One area of IR that has seen a great deal of interest and has had many exciting advances made in it, is CLIR. The goal of CLIR is to allow users to make queries in one language and retrieve documents in one or more other languages. The resulting documents can then be translated into the language used for the query to allow the user to get the gist about the information retrieved. For example, a user makes a query in English about “flower arrangement” and receives documents back in Japanese about “Ikebana” which is Japanese flower arrangement.

Most systems in CLIR use some type of translation. While there exist non-translation methods, such as cognate matching [2], latent semantic indexing [10], and relevance models [4], here the predominate method is still translation. As such one of the main problems in CLIR is dealing with

language translation. What should be translated, how should it be translated, and how to eliminate bad translations are some of the major areas of research in CLIR. In addition how to acquire large enough amounts of translation data is also an active topic for research.

3.1 CLIR survey for Indian languages and Foreign Languages

In [11], the task is to retrieve relevant documents from an English corpus in response to a query expressed in different Indian languages including Hindi, Tamil, Telugu, Bengali and Marathi. A word alignment table have been used that was learnt by a Statistical Machine Translation (SMT) system trained on aligned parallel sentences, to map a query in source language into an equivalent query in the language of the target document collection. The relevant documents are then retrieved using a Language Modeling based retrieval algorithm. This work has been tested on CLEF 2007 data set.

The most commonly used vocabulary in Indian language documents found on the web contains a number of words that have Sanskrit, Persian or English origin. [12] Used approximate string matching techniques to exploit relatively large number of cognates among Indian languages, which are higher when compared to an Indian language and a non-Indian language. An approach to identify cognates was presented and make use of them for improving dictionary based CLIR when the query and documents both belong to two different Indian languages. Experiments using a Hindi document collection and a set of Telugu queries were conducted and report the improvement due to cognate recognition and translation.

The main objective of this work [13] is to analyze and evaluate the retrieval effectiveness of various indexing and search strategies based on test collections written in four different languages: English, French, German, and Italian**.** Data base merging strategies have been used. Experiments are done in CLEF 2000 corpora.

This paper [14] describes an approach that employs user-assisted query translation to help searchers better understand the system's operation.[15] This article identifies the key issues in dictionary-based CLIR, develops unified frameworks for term selection and term translation that help to explain the relationships among existing techniques, and illustrates the effect of those techniques using four contrasting languages for systematic experiments with uniform query translation architecture. The other works are given in table 1.

The table I describe the foreign languages which are involved in the CLIR/MLIR system, the translation technique/method and finally the evaluation initiatives used in the research work for experiments are enumerated. The table II describes the Indian languages which are involved in the CLIR/MLIR system, the translation technique/method and finally the evaluation initiatives used in the research work for experiments are enumerated.

4. MLIR

MLIR facilitates the uses of queries in one language to access documents in different languages. In recent years, large amount of multilingual information is created and disseminated. Due to this reason it attracts the attention of the researchers lately. In order to retrieve this multilingual information efficiently the adaptation of traditional IR systems has been considered. That is query translation and document translations have been used. The problem of MLIR is an extension of the general problem of monolingual information retrieval [16].

Table 1. An overview of CLIR and MLIR research works in foreign languages.

Authors	Languages	Method/Technique	Evaluation Initiatives
Fujii, A., Ishikawa, T., 2001[19]	J to E and E to J	Query translation and Document translation	NTCIR -2 Collection
Jialun Qin, Yilu Zhou, Michael Chau, and Hsinchun Chen, 2003 [22]	E to Ch	Dictionary based query translation	TREC Collection
David A. Hull, Gregory Grefenstette, 1996 [29]	E to F	Dictionary based query translation	Documents Collection
Chen-Yu Su, Tien-Chien Lin and Shih-Hung Wu, 2007 [30]	Ch to J and K	Dictionary based query translation	NTCIR -6
Paraic Sheridan, Jean Paul Ballerini, 1996 [31]	G to I	Thesaurus-based query expansion	Documents Collection
Wen-Cheng Lin, Hsin-Hsi Chen, 2003 [20]	J to E and Ch	Query translation	NTCIR-3
David A. Hull , Gregory Grefenstette, 1996 [32]	E to F	Dictionary based query translation	Documents Collection
Peter A. Chew and Ahmed Abdelali, 2008 [33]	E, R, S, F and A	Latent Semantic Analysis	Bible and Quran data
Su Liu, 2001 [34]	E to Ch	Dictionary based query translation	TREC Collection
Mizera-Pietraszko J, 2009 [35]	E to F and F to E	Meta data search	Documents Collection
Turdi Tohti, Winira Musajan, Askar Hamdulla, 2008 [36]	Uyghur, Kazak, Kyrgyz	Query phase reconstruction, character coding	Website data
Marshall Ramsey, Thian-Huat Ong, Hsinchun Chen, 1998 [37]	Ch and J	Dictionary-lookup, phonetic, radical, and mnemonic	Training data
Dong-Mo Zhang, Huan-Ye Sheng, Fang Li and Tian-Fang Yao, 2002 [38]	E, G and Ch	Case based reasoning and machine learning	Documents

Kazuyuki Yoshinaga, Takao Terano, Ning Zhong, 1999 [39]	J and E	Web Information Collector, Document classifier, Ontology generator and Search engine	Web documents
Hsin-Chang Yang, Chung-Hong Lee, 2008 [45]	E and Ch	Parallel corpora	Bilingual corpus documents
Hassina Aliane, 2006 [40]	A, F and E	Ontology based Approach (corpora)	Trilingual corpus documents
Chung-hsin Lin and Hsinchun Chen, 1996 [41]	Ch and E	Indexing and Classification approach	Multilingual Databases
Jeffrey A. Rydberg-Cox, Lara Vetter, Stefan M. Rüger, Daniel Heesch [42]	Greek, Latin and Old Norse	Query translation	Search engine results
Shuang-Qing Yuan, Fang Li, and Huan-Ye Sheng, 2002 [43]	Ch and E	Novel approach for finding terminology translations from hyperlinks	Website Links (parallel or unparallel corpus)
Akiko Aizawa, 2002. [44]	E and J	Evolutionary framework	NTCIR-J1

Table 2. An Overview of CLIR and MLIR Research Works In Indian Languages

Authors	**Languages**	**Method/Technique**	**Evaluation Initiatives**
Jagadeesh, J. and Kumaran, K , 2007 [11]	(H, Ta, Te, Be, Ma and E) to E	structural query translation, Language Modeling based retrieval algorithm	CLEF 2007
Ranbeer Makin, Nikita Pandey, Prasad Pingali, and Vasudeva Varma,2007 [12]	Te to E	Dictionary based query translation, String matching	Documents Collection
A. Kumaran, Jayant and R. Haritsa, 2005 [23]	E to (Ta and F)	Semantic text matching	Standard SQL:1999 Constructs
Prasad Pingali and Vasudeva Varma, 2006 [24]	(H and Te) to E	Dictionary based query translation, Lucene search engine, vector based ranking model	CLEF 2006

Tune, K. K, Pingali, P., Varma, V., 2007 [25]	Oromo to E	Dictionary based query translation, Approximate string matching	CLEF 2006
Sethuramalingam S and Vasudeva Varma, 2008 [26]	E to H and H to E	Dictionary based query translation, Mapping based approach for transliteration, Lucene's BM25 algorithm for ranking	FIRE-2008
Manoj kumar Chinnakotla and Om P.Damani, 2009 [27]	E to (H, Te, Ta)	Machine Transliteration	NEWS 2009
Anurag Seetha, Sujoy Dos and M. Kumar, 2007 [28]	E to H	Dictionary based query translation	Documents Collection
P. Sujatha, P. Dhavachelvan, V. Narasimhulu 2010 [46]	E to (Te and Ta)	Dictionary based query translation	Documents Collection

MLIR can be thought of as a combination of machine translation and traditional monolingual information retrieval. Most research has focused on locating and exploiting translation resources with which the user's search requests or target documents (or both) are translated into the same language.

A multilingual data collection is a set of documents that are written in different languages. There are two types of multilingual data collection. The first one contains several monolingual document collections. The second one consists of multilingual-documents. A multilingual-document is written in more than two languages. Some multilingual-documents have a major language, i.e. most part of the document is written in the same language.

4.1 MLIR Survey for Indian Languages and Foreign Languages

There are no works in Indian languages as such and very few works are available for foreign languages. They are presented as follows: Most systems in MLIR use some type of translation. While there exist non-translation methods, such as: Translation-free technique is based on an ontological representation of documents and queries. A multilingual ontology for documents/queries representation has been used [18].

Integrate query and document translation with monolingual retrieval to improve retrieval accuracy have been presented in [19], and perform clustering to improve browsing efficiency. Finally, an entropy-driven technique in evaluating clustering methods has been introduced.

Participated in NTCIR-2 Japanese/English cross-language (J-E and E-J) and multi-lingual (J-JE and E-JE) information retrieval tasks. In this paper, performance evaluation is done with respect to the NTCIR-2 collection. The paper [20] deals with Chinese, English and Japanese MLIR. Merging problem in distributed MLIR is studied. [21] Presented a MLIR based on knowledge representation model. This model permits to describe the semantic of document in a multilingual context. This model, called semantic graph, is an extension of the Sowa's model of conceptual graphs where different vocabularies' are available. [22] Developed and evaluated a multilingual English-Chinese Web portal in the business domain. A dictionary-based approach has been adopted that combines phrasal translation, co-occurrence analysis, and pre- and post-translation query expansion.

Recently, a number of tracks and workshops have sprung up to support research in this area. They are TREC (Text Retrieval Conference), came up in 2008, its first conference held in USA and it is sponsored by the National Institute for Standards and Technology (NIST). It has organized 26 tracks: they are Question answering track, Genomics track, HARD track, robust retrieval track, Terabyte track, etc. CLEF (Cross Language Evaluation Forum) first workshop held in Europe and closely followed the TREC model. Its motivation is to develop linguistic resources and retrieval in each languageand there is another new track ImageCLEF which combines access to textual and graphic data. NTCIR (NII Test Collection for IR Systems) Project is a yearly competition in Japan that covers many topics including CLIR. The workshops have started since 1997and it has apatent, a Web and a question-answering track.

5. Conclusion

Cross-lingual and Multi-lingual IR provides new paradigms in searching documents through myriad varieties of languages across the world and it can be the baseline for searching not only among two languages but also in multiple. This report explains a description on cross-lingual and multi-lingual IR, its challenges and current methods, techniques and evaluation tracks to overcome problems for efficient and resourceful searching. This report meant for reviewing not all but some of the latest researches in the area of cross-lingual and multi-lingual IR.

References

[1] Abusalah, M., J. Tait, and M. Oakes: Literature Review of Cross Language Information Retrieval. In: World Academy of Science, Engineering and Technology 4, 175-177 (2005).

[2] Buckley, C., M. Mitra, J. A. Walz and C. Cardie.: Using clustering and super concepts within SMART. TREC 6, Information Processing and Management 36 (2000), pp. 109–131.

[3] Korfhage, R. R.: Information Storage and Retrieval. In: John Wiley and Sons, 1997.

[4] Lavrenko, V., M. Choquette and W. B. Croft.: Cross-lingual relevance models. In: SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (2002), pp. 175–182.

[5] McNamee, P. and J. Mayfield.: Comparing cross-language query expansion techniques by degrading translation resources. In: SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (2002), pp. 159–166.

[6] Zhang, Y. and P. Vines.: Using the web for automated translation extraction in cross-language information retrieval. In:SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (2004), pp. 162–169.

[7] Belkin, N.J., Dumais, S.T., Scholtz, J. & Wilkinson R.: Evaluating interactive information retrieval systems: opportunities and challenges.In: CHI Extended Abstracts 2004: 1594-1595.

[8] Peters, C. and Braschler, M.: Cross Language System Evaluation: The CLEF Campaigns.In: Journal of the American Soc. for Inf. Sci. and Tech. Vol. 52(12) (2001) 1067-1072.

[9] Voorhees, E.M. & Harman, D.: Overview of TREC 2001.In: NIST Special Publication 500-250: Proceedings of TREC2001, NIST, 2001.

[10] Dumais, S. T., T. A. Letsche, M. L. Littman and T. K. Landauer.: Automatic cross-language retrieval using latent semantic indexing.In: AAAI Spring Symposuim on Cross-Language Text and Speech Retrieval, 1997.

[11] Jagadeesh, J. and Kumaran, K.: Cross-Lingual Information Retrieval System for Indian Languages.In: Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, pages: 80-87.

[12] Ranbeer Makin, Nikita Pandey, Prasad Pingali, VasudevaVarma.: Approximate String Matching Techniques for Effective CLIR Among Indian Languages. In: WILF 2007: 430-437

[13] Jacques Savoy.: Cross-language information retrieval: experiments based on CLEF 2000 corpora. In: Information Processing and Management 39 (2003) 75–115.

[14] Douglas W. Oard et al.: User-assisted query translation for interactive cross-language information retrieval. In: Information Processing and Management 44 (2008) 181–211.

[15] Gina-Anne Levow et al.: Dictionary-based techniques for cross-language information retrieval. In: Information Processing and Management 41 (2005) 523–547

[16] Ahmed Abdelali, James R. Cowie, David Farwell, William C. Ogden.: UCLIR: a Multilingual Information Retrieval Tool. In: Inteligencia Artificial, RevistaIberoamericana de Inteligencia Artificial 22: 103-110 (2003).

[17] P. Clough, J. Gonzalo, J. Karlgren, E. Barker, J. Artiles, V. Peinado.: Large-Scale Interactive Evaluation of Multilingual Information Access Systems - the iCLEF Flickr Challenge. In: 30th European Conference on Information Retrieval (ECIR 2008), pp. 33-38, 2008.

[18] HassinaAliane.: An Ontology Based Approach to Multilingual Information Retrieval. In: Artificial Intelligence Division, Research Center on Scientific and Technical Information, 2006

[19] Fujii, A., Ishikawa, T.: Evaluating Multi-lingual Information Retrieval and Clustering at ULIS. In: Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization. Tokyo. Japan. March 2001. P.5-144-5-148

[20] Wen-Cheng Lin, Hsin-HsiChen.:Description of NTU Approach to NTCIR3 Multilingual Information Retrieval. In: Proceedings of the third NTCIR workshop on research in information retrieval automatic text summarization and question answering. Tokyo: National Institute of Informatics, 2003

[21] Catherine Roussey et al.: SyDoM: A Multilingual Information Retrieval System for Digital Libraries.In: Electronics Publishing '01 – 2001 in the Digital Publishing Odyssey, IOS Press, 2001.

[22] Jialun Qin, Yilu Zhou, Michael Chau, Hsinchun Chen.: Supporting Multilingual Information Retrieval in Web Applications: An English-Chinese Web Portal Experiment. In: ICADL 2003, pp. 149-152.

[23] A. Kumaran, Jayant R. Haritsa: SemEQUAL: Multilingual Semantic Matching in Relational Systems. In: DASFAA 2005, pp. 214-225

[24] Prasad Pingali and VasudevaVarma.: Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006. In: September, Alicante, Spain.

[25] Pingali, P., Varma, V., Tune, K.K.: Evaluation of Oromo-English Cross-Language. Information Retrieval. In: IJCAI 2007 Workshop on CLIA, Hyderabad, India

[26] Sethuramalingam S and VasudevaVarma.: IIIT Hyderabad's CLIR experiments for FIRE-2008. In: The working notes of First Workshop of Forum for Information Retrieval Evaluation (FIRE) 2008 Kolkata.

[27] ManojkumarChinnakotla and Om P.Damani.: Experiences with English-Hindi, English-Tamil and English-KannadaTransliteration Tasks at NEWS 2009. In: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration, pp: 44-47, 2009.

[28] AnuragSeetha, Sujoy Dos and M. Kumar.: Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method.In: Proceedings of the 10th International Conference on Information Technology, 2007

[29] David A. Hull, Gregory Grefenstette: Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval.In: SIGIR 1996: 49-57

[30] Chen-Yu Su, Tien-Chien Lin and Shih-Hung Wu.: Using Wikipedia to Translate OOV Terms on MLIR.In: Proceedings of NTCIR-6 Workshop Meeting, May 15-18, 2007, Tokyo, Japan.

[31] Paraic Sheridan, Jean Paul Ballerini: Experiments in Multilingual Information Retrieval Using the SPIDER System. In: SIGIR 1996: 58-65

[32] David A. Hull , Gregory Grefenstette: Experiments in Multilingual Information Retrieval.In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996

[33] Peter A. Chew and AhmedAbdelali.: The Effects of Language Relatedness on Multilingual Information Re-trieval: A Case Study With Indo-European and Semitic Languages. In: The second International workshop on cross lingual information access, 2008.

[34] Su Liu.: ECIRS: an English-Chinese Cross-language Information-retrieval System. In: IEEE International Conference on Systems, Man, and Cybernetics, Volume 2, 2001.

[35] Mizera-Pietraszko J.:Interactive Document Retrieval from Multilingual Digital Repositories. In: IEEE Xplore Digital Library, IEEE Computer Society Press, str. ICADIWT 2009,pp-423-428, 2009.

[36] TurdiTohti, WiniraMusajan, AskarHamdulla: Character Code Conversion and Misspelled Word Processing in Uyghur, Kazak, Kyrgyz Multilingual Information Retrieval System. In: ALPIT 2008: 139-144

[37] Marshall Ramsey, Thian-HuatOng, Hsinchun Chen: Multilingual Input System for the Web - An Open Multimedia Approach of Keyboard and Handwriting Recognition for Chinese and Japanese. In: ADL 1998: 188-194.

[38] Dong-Mo Zhang, Huan-Ye Sheng, Fang Li and Tian-Fang Yao.: The model and design of a case-based reasoning multilingual natural language interface for database. In: Proceedings of the first international conference on machine learning and cybermetics,2002.

[39] Kazuyuki Yoshinaga, Takao Terano, NingZhong.: Multi-lingual Intelligent Information Retriever with Automated Ontology Generator.In: Third International Conference on Knowledge based Intelligent Information Engineering Systems, 1999.

[40] Hsin-Chang Yang, Chung-Hong Lee: Multilingual Information Retrieval Using GHSOM.In: ISDA (1) 2008: 225-228

[41] Chung-hsinLin andHsinchun Chen.: An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual (Chinese-English) Documents.In: IEEETransactions on Systems, Man, And Cybernetics-Part B: Cybernetics, Vol. 26, No. 1, February 1996

[42] Jeffrey A. Rydberg-Cox, Lara Vetter, Stefan M. Rüger, Daniel Heesch: Cross-lingual searching and visualization for greek and latin and old norse texts. In: JCDL 2004: 383

[43] Shuang-Qing Yuan, Fang Li, and Huan- Ye Sheng.: Finding terminology translations from hyperlinks on the internet.In: Proceedings of the first international conference on machine learning and cybermetics, 2002

[44] Akiko Aizawa.: A co-evolutionary framework for clustering in information retrieval systems. In: National Institute of Informatics, 2002.

[45] Hsin-Chang Yang, Chung-Hong Lee: Multilingual Information Retrieval Using GHSOM. In: ISDA (1) 2008: 225-228

[46] P. Sujatha**,** P. Dhavachelvan, V. Narasimhulu, "Evaluation of English-Telugu and English-Tamil Cross Language Information Retrieval System using Dictionary Based Query Translation Method", International Journal of Computer Science and Information Security (IJCSIS), Volume: 8 Issue: 2, Year: May 2010, Pages: 314-319.