State of Art: Cross Lingual Information Retrieval System for Indian Languages

A. Nagarathinam
Research Scholar, Manonmaniam Sundaranar University &
Associate Professor, AVC College of Engineering
Tamilnadu-609305, India

Dr. S. Saraswathi
Associate Professor
Department of Information Technology
Pondicherry Engineering College,
Puducherry-605004, India

ABSTRACT
The number of Web Users accessing the Internet becomes huge now a day. Also, all sorts of information can be obtained anytime by anybody from the Web. In India plenty of people are speaking diversified local languages. Only a very few percentage of population know English language and they can express their queries in English in a right way. Though, the network shrank the Globe, the language diversification is a great barrier to enjoy the benefits of the digital life. Cross Lingual Information Retrieval provides the solution for that barrier, by allowing the user to ask the query in the local language and then to get the documents in another language (English).

In Cross Lingual Information Retrieval environment, the data interaction is not restricted by the language. Information Retrieval tries to match a user's description of their information need with relevant information in a collection of documents or other data. Thereby it tries to resolve the language mismatch between the documents and queries.

Keywords
Monolingual IR, CLIR, Machine Translation, Bilingual Dictionary, Corpus

1. INTRODUCTION
Information Retrieval can be termed as a reasoning process that is used to recognize the existence of relationship between documents and queries and also to assess how strong the relationship is [1].

1.1 Types of Information Retrieval
Monolingual Information Retrieval System - refers to the Information Retrieval system that can identify the relevant documents in the same language as the query was expressed. *Cross Lingual Information Retrieval System* (CLIR) is a sub field of Information Retrieval dealing with retrieving information written in a language different from the language of the user's query.

1.2 Need for Cross Lingual Information Retrieval
A survey on Online Computer Library Center states that English is still the dominant language in the Web [2]. Global Internet Usage statistics shows that number of web access by the non-English users is tremendously increased. But, all of them are not able to express their query in English [3].

In India, students learn more than one language from their childhood and more than 30% of the population can read and understand Hindi apart from their native language [4]. We also have the controversy opinion [5] such that the information in another language will be useful unless users are already fluent in that language.

Also, the Precision and the Recall of the content provided by the Tamil search engines or by the improper query are low in performance [6].

1.3 Types of Translations in CLIR
There are two types of translations namely Query translation and Document Translation.

In Query Translation, the given query will be converted from Native language to English and will search the database to get the documents in English. Later the retrieved documents in English language can be converted to Native language.

In Document Translation, all the documents are translated from English to Native language. It allows the user to ask query in Native language and now the searching will take place to obtain the resultant documents in Native language.

Among the two, the former is easier compared with later, because of the size of translation. The efficiency of the query translation depends on the best translation words and weight for the given query.

But, the drawback with Query Translation is the given query normally will be short and hence ambiguity problem may arise. Since, Document Translation is not feasible, in most of the research works, Query Translation will be carried out instead of Document Translation.

2. RESEARCH TOWARDS CLIR
2.1 Text Retrieval Conference
The Text Retrieval Conference (TREC), a conference series co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense, was started in 1992 with the aim of encouraging research in text retrieval by using realistically large test collections [7].

2.2 Cross Language Evaluation Forum
CLEF evolved out of a Cross-Lingual Information Retrieval (IR) track involving European languages that was organized as a part of TREC during 1997–1999.

2.3 NII Test Collection for IR Systems
NTCIR was evolved in 1996 for Chinese language track and Cross-lingual English-Chinese track in 2000, specifically for the East Asian languages such as Chinese, Japanese, and Korean.

2.4 Forum for Information Retrieval Evaluation

The first evaluation exercise conducted by FIRE was completed in 2008 by covering the four languages Bengali, Hindi, Marathi, and English. The suitable Corpora in terms of size, genre, and time period covered was taken as the domain for their experiments.

2.5 CLIA Consortium

The "Development of Cross Lingual Information Access (CLIA) System" is the project funded by the Government of India, Ministry of Communications & Information Technology and Department of Information Technology. The CLIA Consortium includes 11 Institutions to implement this project namely IIT Bombay, IIT Kharagpur, IIT Hyderabad, AU-KBC Chennai, AU-CEG Chennai, ISI Kolkata, Jadavpur University Kolkata, C-DAC Pune, C-DAC Noida, Utkal University Bhubaneswar and STDC, DIT New Delhi.

The objective of this Consortium is to create a Portal that enables any user to give a query in one Indian language, can access the documents available in three possibilities such as accessing the documents in the language of the query, in Hindi (if the query language is not Hindi) and in English.

3. APPROACHES USED IN FIRE EXPERIMENTS

There were 9 Institutes participated in FIRE 2008 namely, AU-KBC, IIT Hyderabad, IIT Bombay(1), IIT Bombay (2), ISI Kolkata, Johns Hopkins U., Microsoft Research, U. Maryland and U. Neuchatel.

3.1 Hindi- Monolingual Runs

Mcnamee P [8], used the HAIRCUT system and obtained the highest MAP score using character 5-grams and Pseudo Relevance Feedback. The language modeling approach proposed by Hiemstra [9] was used to compute document similarity scores.

Dolamic L., Savoy J [10], had created a stopword list for the most frequently occurring words in the corpus. **Sparck Jones K., Walker S., and Robertson S**[11]**,** used character 4-grams, a stemmer, pseudo-relevance feedback, DFR model and BM25 formula for Ranking.

Paik J. H., Parui S. K [12], tested a variant of the DFR model (IFB2). **Sethuramalingam S., Varma V** [13], used Lucene and BM25 term-weighting scheme.

3.2 Hindi Cross-Lingual Runs

U. Maryland (UMD) obtained the highest MAP scores by using Translation Matching (DAMM) technique and PRF before translation.

Sethuramalingam S., Varma V [13], used a bilingual lexicon for query translation and a Conditional Random Field-based Named Entity Recognizer. The lexicon was constructed from Shabdanjali[14], the Hindi WordNet [15] and manually constructed Hindi-English dictionary.

Padariya N., Chinnakotla M., Nagesh A., Damani O. P. [16], used a Rule-based transliteration method.

3.3 Bengali

Dolamic L., Savoy J. [10], used a light stemmer and a Z -score-based fusion method and obtained the best MAP among all Bengali monolingual submissions.

Johns Hopkins U (JHU) compared various character n-gram tokenization schemes for Bengali.

3.4 Marathi

Dolamic L., Savoy J. [10], used a light stemmer with character 4-gram indexing.

Johns Hopkins U (JHU) used character n-gram tokenization for all runs.

3.5 English

Rao P.R. And Sobha L [17] submitted two runs using Tamil as the query language with ontology-based query expansion. Bilingual dictionary was used to translate the Tamil query. OOVs and named-entities were transliterated using a statistical method. They achieved 86.5% of the MAP obtained by the best monolingual English run.

Udupa R., Jagarlamudi J., Saravanan K. [18], used a probabilistic translation lexicon, Transliteration Mining technique and a language modeling approach.

3.6 Results Obtained in FIRE Experiments

In the Cross Language Information Retrieval, all the groups attempted to translate the query from the given language to English language. The root words were identified using Porter Stemmer method. Also, the documents were indexed with Lucene Indexer. The results show that the techniques such as effective term-weighting and pseudo relevance feedback yielded the reasonable results for monolingual retrieval. A language-independent, character n-gram-based indexing method works well for the Hindi and gives the better MAP for Monolingual Hindi run 0.3487. U. Maryland (UMD) obtained the highest MAP scores 0.2793, with their best run achieving about 80% of the performance of the best Hindi monolingual run.

Table 1 shows the result of Cross Lingual run to English track.

4. LITERATURE REVIEW

Pattabhi R.K Rao T and Sobha. L [19], proposed Cross Lingual Information Retrieval between Tamil and English languages by using translation, transliteration and query expansion. A Tamil – English bilingual dictionary, which is of 150K size, was used for the translation of the query and statistical method using n-grams based approach was referred for the transliteration.

They worked with the domain, "The Telegraph", an English magazine in India that consists of 125638 documents with 50 topics in Tamil. The query asked in Tamil had been converted into English language. WordNet was used to provide more synonyms. In order to get the more relevant document for the given query, Okapi BM25 Ranking with a boost factor was used.

Jagadeesh Jagarlamudi and A Kumaran [20], proposed a Cross-Lingual Information Retrieval System for Indian Languages in CLEF 2007. The documents were provided in English and queries were given in non-English language including Hindi, Telugu, Bengali, Marathi and Tamil. The system had 1000 relevant documents. They organized the magazine "Los Angeles Times" as the domain which includes 1,35,153 English news articles consisting of 50 topics with 1000 relevant documents collected from 2002.

Table 1. Results for the X to English Cross-Lingual Task

Group	Indexing Units	Translation resource	Term weighting	Query expansion	MAP
Best English monolingual run					0.5572
AU-KBC	words + morphological analyzer	Wordnet, ontology	BM25	Ontology-based	0.4821
MSR	words	probabilistic lexicon + transliteration mining	LM (PC)	No	0.4526
MSR	words	probabilistic lexicon + transliteration mining	LM (PC)	No	0.4495
MSR	words	probabilistic lexicon + transliteration mining	LM (PC)	No	0.4261
MSR	words	probabilistic lexicon + transliteration mining	LM (PC)	No	0.4140

They applied Language Modeling based retrieval algorithm, Stop word removal and Porter Stemmer algorithms. Structural Query translation approach along with a Bilingual statistical dictionary having Hindi to English word alignment of size 100K was used for the translation of query into English language.

Word by word translation with threshold probability was applied to translate the query in Hindi language into English. The performance of CLIR system was 73% of monolingual system.

S. Thenmozhi and C. Aravindan[21], provided a Tamil – English Cross Lingual Information Retrieval System for Agriculture Society. They used morphological analyzer to get the base form of the root words in the query. Machine Translation approach with Bi-lingual Dictionary of 5.08MB size was developed for translation process. Part of Speech tagging and WordNet was used to retrieve words. The Mean Average Precision for this approach was 95%.

Saraswathi et al[22], proposed a Bilingual Information Retrieval System for English and Tamil for the Festival domain. Documents related to Christmas, New Year, Easter and Good Friday were collected for English language and documents related to Diwali, Pongal, Navarathiri, Kaarthigai Deepam and Vinayagar chathurthi were collected for Tamil language. Total of 200 documents were collected for both the languages.

They applied POS Tagger, Machine translation method and Ontology tree. It was found that the relevance of information retrieved was improved by 40% for English and 60% for Tamil language.

Karush Arora et al[23], provided a Cross Lingual Information Retrieval efficiency improvement through transliteration. They took tourism as the domain and used a Bilingual Dictionary having Punjabi and Hindi documents. In tourism domain, it has been observed that most of the queries contain 36-43% of OOV words and 90% of the OOV words are proper noun such as name of a place and monuments. They used rule based transliteration system. They concluded that, Hindi to Punjabi transliteration results is better than Punjabi to Hindi results due to the lesser number of consonant conjuncts in Punjabi than in Hindi.

Chaware and Srikantha Rao[24], projected Domain specific Information Retrieval in Multilingual environment by considering a shopping mall as the domain. The user can pose the query in Hindi, Marathi or Guajarati and the back end data is stored in English. Using Character-by-Character mapping the query will be converted to English. When there was an exact match found, using translation module the keywords were converted to local language by considering Character-to-ASCII mapping. They concluded that the efficiency of the Information retrieval depends on the minimum number of keys to be mapped to enter a local string.

Saravanan et al[25], found that both by generating transliterations directly or transitively and mining possible transliteration equivalents from the documents retrieved in the first pass, the efficiency of cross lingual information retrieval will be improved.

Table 2 shows the Analysis of Literature Review.

5. PROCESS INVOLVED IN IR SYSTEMS

5.1 Machine Translation

Machine Translation [26] explores the use of computer software to translate text or speech from one natural language to another. Google Translate Machine is one of translation system which is used to translate the solution obtained.

5.2 Bi Lingual Dictionary

Dictionary approach is used to translate the Query. CLIR depends on the quality and coverage of the dictionary. The Quality refers to the ability of the dictionary to provide an precise translation of a query. Coverage refers to the ability of the bilingual dictionary to provide translations for a wide range of words [27].

Manually created bilingual dictionary provides the good quality and poor coverage.

5.3 Corpus

In linguistics, a corpus (plural corpora) or text corpus [28] is a large and structured set of texts. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe.

A corpus may contain texts in a single language (Monolingual Corpus) or text data in multiple languages (Multilingual Corpus).

Table 2 - Analysis of Literature Review.

Author(s)	Language -Query	Document Lang	Domain	Indexing Unit	Translation	Transliteration	Query Expansion	Ranking	Result for CLIR
Pattabhi R.K Rao T and Sobha. L	Tamil and English	English	The Telegraph	Tamil Morphological Analyzer, Lucene Indexer & *Porter Stemmer*	Bilingual Dictionary	Statistical method	WordNet & Description Field	*Okapi BM25*	MAP: 0.3980 Recall Precision : 0.3742
Jagadeesh Jagarlamudi and A Kumaran	Hindi, Tamil, Telugu, Bengali and Marathi	English	Los Angeles Times	Stop word removal and Porter Stemmer	Bilingual statistical dictionary & Word by word translation	Yes	No	LM	CLIR Performance: 73% of monolingual system
S. Thenmozhi and C. Aravindan	Tamil and English	English	Agriculture	Morphological Analyzer, Pos Tagger	Machine Translation, Bi-lingual Dictionary & Local Word Reordering	Named Entity Recognizer	WordNet	No	Mean Average Precision : 95% of Monolingual system
Saraswathi et al	Tamil and English	Tamil and English	Festival	Tamil Morphological Analyzer, Pos Tagger, Porter Stemmer	Machine Translation	No	Ontology	Page Rank	Tamil - Increased by 60% English - Increased by 40%
Karush Arora et al	Punjabi and Hindi	Punjabi and Hindi	Tourism	Porter Stemmer	Bilingual Dictionary	Rule Based Transliteration	No	-	Hindi to Punjabi transliteration results are better than Punjabi to Hindi results
Chaware and Srikantha Rao	Hindi, Guajarati, Marathi	English	Shopping Mall	Text-to-phonetic algorithm	No	Char by Char, Char to ASCII mapping	No	-	Efficiency depends on minimum number of keys to be mapped
Saravanan et al	Tamil , English, Hindi	English	The Telegraph	Porter Stemmer & Alignment Model	Probabilistic Lexicon & Parallel Corpora & Bilingual dictionary	Machine Transliteration, Transliteration Similarity Model	No	LM	MAP Monolingual IR: 0.5133 Hindi to Eng: 0.4977 & Tamil to Eng: 0.4145

- The tagged corpus is that which is tagged for part-of-speech.
- A parallel corpus contains texts and translations in each of the languages involved in it.
- Aligned corpus is a kind of bilingual corpus where text samples of one language and their translations into other language are aligned, sentence by sentence, phrase by phrase, word by word, or even character by character.

5.4 Treebank

A Treebank or parsed corpus is a text corpus in which each sentence has been parsed, i.e. annotated with syntactic structure. Syntactic structure is normally represented as a tree structure.

5.5 Named Entity Recognizer

Named entity recognition (NER) is a subtask of information extraction that looks for to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, etc.

5.6 Out-of-Vocabulary Terms

Any term not found in a dictionary normally termed as OOV. It can be either a noun phrase or a noun term. Transliteration represents a string matching process that works best when the two languages having a shared common alphabet. Phonetic mapping is needed for the languages those not having similar alphabets.

5.7 Wikipedia

Wikipedia is the free online encyclopedia used in CLIR. Because of its scalability nature and provider for up to date information normally it is used as Named Entity in Bilingual Dictionaries. It permits to transliterate the phrases as well as names.

5.8 Morphological analyzer

Morphological Analyzer is a software component capable of detecting morphemes in a piece of text. Atcharam - Tamil Morphological Analyzer uses a dictionary of 20000 root words based on fifteen categories. It has noun and verb analyzer based on 125 rules.

5.9 PoS Tag the sequence.

Part-of-speech tagging is the process of assigning a part-of-speech like noun, verb, pronoun, preposition, adverb, adjective or other lexical class marker to each word in a sentence.

5.10 Word Sense Disambiguation

The process of identifying the correct sense for a given word sequence is known as Word Sense Disambiguation [29].

Polysemy [30] is a lexeme having more than one sense. E.g. Crane can be a machine or a bird.

Homonyms are words that are spelled the same and have different meanings. E.g. bank

5.11 Precision

Precision refers the number of retrieved relevant concepts judged as relevant by the subjects over the total number of retrieved concepts [22].

5.12 Recall

Recall refers the number of retrieved concepts judged as relevant by the subjects over the number of relevant concepts judged and suggested by the subject.

Precision and Recall are used to evaluate the effectiveness of the CLIR System.

Table 3 summarizes the features of Existing techniques used in CLIR.

Table 3 - Analysis of Existing Techniques in CLIR

Machine Translation	Uses the computer software to translate text or speech from one natural language to another.
BiLingual Dictionary	Dictionary approach is used to translate the Query from one language to another language.
Corpus	Used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe.
Tagged corpus	Used in part-of-speech.
Parallel corpus	Contains texts and translations in each of the languages involved in it.
Aligned corpus	Text samples are aligned, sentence by sentence, phrase by phrase, word by word, or even character by character
Treebank	A text corpus in which each sentence has been parsed into Tree structure
Named Entity Recognizer	Used for Transliteration
Wikipedia	The free online encyclopedia used as Named Entity in Bilingual dictionaries.
Morphological Analyzer	A software component capable of detecting morphemes in a piece of text.
Atcharam - Tamil Morphological Analyzer	It uses a dictionary of 20000 root words based on fifteen categories. It has noun and verb analyzer based on 125 rules.
PoS Tagger	Part-of-speech tagging is the process of assigning a part-of-speech like noun, verb, pronoun, preposition, adverb, adjective or other lexical class marker to each word in a sentence
Word Sense Disambiguation	The process of identifying the correct sense for a given word sequence.
Polysemy	A lexeme having more than one sense.
Homonyms	Words those are spelled same and have different meanings

6. Subsystems of CLIR system

6.1 Input processing Sub System:

Input processing system analyses the query entered by the user using language processing tools. More relevant terms can be added to expand the query and based on its analysis either it translates or transliterates all the query terms to the target language and provides this as input to the search module.

6.2 Search

Search sub system is used to crawl the web and download files for a specific language and domain. Then the text part in these documents can be extracted to perform certain processing. Page ranking algorithm is used to arrange all the documents to identify the most useful document

6.3 Processing of Retrieved Documents

Processing sub system generates the output document relevant to the given query by considering the *Similarity of terms and Ranking.*

Figure 1 shows the major elements of the CLIR system.

7. CONCLUSION

The World's top 25 most widely spoken languages include the Indian languages Hindi, Telugu, Marathi, Tamil, Gujarati and Bhojpuri. Hence the need for CLIR is increased. This paper presented a survey of some of the aspects of the Cross Lingual Information Retrieval such as types of Information Retrieval, need and types of CLIR, processes involved in CLIR. CLIR removes the linguistic gap between the query submitted by the user and the retrieved document. It allows any user to access the Web and to get the needed information. Also, the Indian local languages usage in Internet will be amplified by the CLIR.

In CLIR, for the query asked in Native language, it searches the English database, and gives the relevant documents in English language. It can be enhanced by giving the results in the language of the Query itself. Also, Similarity Scores can be found for the retrieved documents in both English and Local Language (Query Language) to identify the content of the documents in the two languages. Then, by applying

Summarization and Information Retrieval, the content present more in one language document can be appended with the document in Native language. There by, it is possible to return a document not only in English; but, also in Local language (Query language).

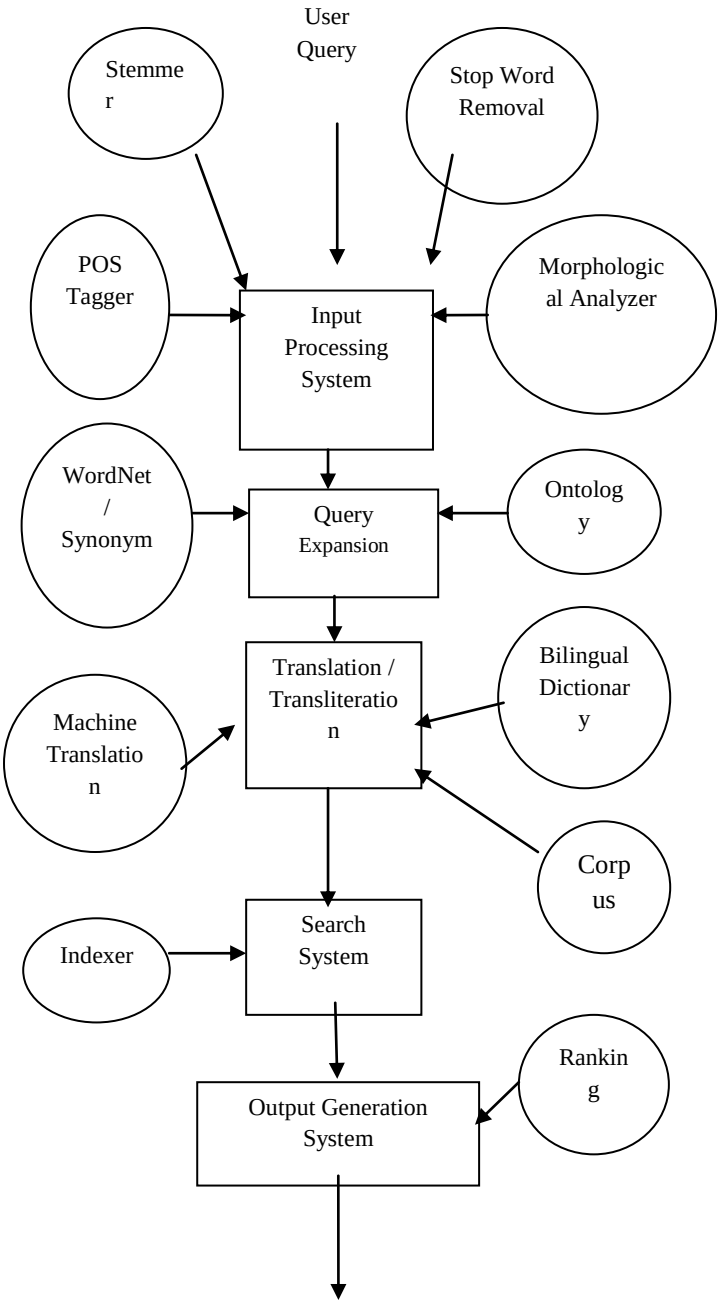

Figure 1 – Major Elements of CLIR

8. REFERENCES

[1] Jianfeng Gao, Jian-Yun Nie, Ming Zhou. Statistical Query Translation Models for Cross-Language Information Retrieval, ACM Transactions on Asian Language Information Processing, 5, 4, December 2006.

[2] Manoj Kumar Cinnakotla, Sagar Ranadive, Pushpak Bhattacharyya and Om Damani, P. Hindi and Marathi to English CLIR at CLEF 2007, Working notes of CLEF 2007.

[3] Internet World Stats: http://www.internetworldstats.com.

[4] Grey Burkhart, E., Seymour Goodman, E., Arun Mehta and Larry Press. The Internet in India: better times ahead?, Journal of ACM Communication, 41, 11, 21-26, 1998.

[5] Isabelle Moulinier & Frank Schilder. What is the future of multi-lingual Information Access?, SIGIR 2006 Workshop on Multilingual Information Access 2006, Seattle, Washington, USA, 2006.

[6] Prasenjit Majumder, Mandar Mitra Swapan Parui and Pushpak Bhattacharyya. Initiative for Indian Language IR Evaluation, Invited paper in EVIA 2007 Online Proceedings, 2007.

[7] Prasenjit Majumder, Mandar Mitra, Dipasree Pal, Ayan Bandyopadhyay, Samaresh Maiti, Sukomal Pal. The FIRE 2008 Evaluation Exercise, ACM Transactions on Asian Language Information Processing, 1-24, September 2010.

[8] Mcnamee, P. N-Gram Tokenization for Indian Language Text Retrieval, Working Notes from Fire 2008, 2008.

[9] Hiemstra, D. Using language models for Information Retrieval, Ph.D. Thesis University of Twente, 2001.

[10] Dolamic, L. and Savoy, J. UniNE at FIRE 2008: Hindi, Bengali, and Marathi IR, Working Notes from Fire 2008, 2008.

[11] Sparck Jones, K., Walker, S., and Robertson, S. A probabilistic model of Information Retrieval: Development and comparative experiment, Journal of Information Processing Management, 36, 6, 779–808, 2000.

[12] Paik, J. H., and Parui, S. K. A simple stemmer for inflectional languages, Working Notes from FIRE 2008 (FIRE'08), 2008.

[13] Sethuramalingam, S., and Varma, V. CLIR experiments for FIRE-2008, Working Notes from FIRE 2008.

[14] About English Hindi dictionary and translation: http://www.shabdkosh.com/shabdanjali.

[15] Hindi Wordnet Online http://www.cfilt.iitb.ac.in/wordnet/webhwn

[16] Padariya, N., Chinnakotla, M., Nagesh, A., and Damani O. P. Evaluation of Hindi to English, Marathi to English, and English to Hindi CLIR at FIRE 2008, Working Notes from FIRE 2008, 2008.

[17] Rao, P. R., and Sobha, L. Submission - Cross lingual information retrieval track: Tamil-English, Working Notes from FIRE 2008, 2008.

[18] Udupa, R., Jagarlamudi, J., and Saravanan, K. Hindi-English cross- language information retrieval, Working Notes from FIRE 2008, 2008.

[19] Pattabhi R. K. Rao., and Sobha, L. Cross Lingual Information Retrieval Track: Tamil – English, Working notes from FIRE 2010, Feb 2010.

[20] Jagadeesh Jagarlamudi and Kumaran, A. Cross-Lingual Information Retrieval System for Indian Languages, Proceedings of CLEF 2007, 2007.

[21] Thenmozhi, D., and Aravindan, C. Tamil-English Cross Lingual Information Retrieval System for Agriculture Society, International Forum for Information Technology in Tamil Conference, October 2009.

[22] Dr. Saraswathi, S., Asma Siddhiqaa, M., Kalaimagal, K., and Kalaiyarasi M. BiLingual Information Retrieval System for English and Tamil, Journal Of Computing, 2,4, 85-89, April 2010.

[23] Karunesh Arora, Ankur Garg, Gour Mohan, Somiram Singla, Chander Mohan. Cross Lingual Information Retrieval Efficiency Improvement through Transliteration, Proceedings of ASCNT 2009, 65-71, 2009.

[24] Chawre, S. M., Srikantha Rao. Domain Specific Information Retrieval in Multilingual Environment, International Journal of Recent Trends in Engineering, 2, 4, 179-181, 2009.

[25] Saravanan, K., Raghavendra Udupa, Kumaran, A. Cross lingual Information Retrieval System Enhanced with Transliteration Generation and Mining, Proceedings of Workshop FIRE 2010, 2010.

[26] Gey, F., He, J., and Chen, A. Manual queries and Machine Translation in cross-language retrieval at TREC-7, in TREC7 Proceedings, NIST Special Publication, 1999.

[27] Bilingual dictionary, available at: http://en.wikipedia.org/wiki/Bilingual_dictionary.

[28] Oren Kurland and Lillian Lee. Corpus structure, language models, and ad hoc Information Retrieval, Proceedings of SIGIR, 194-201, 2004.

[29] Dong Zhou, Mark Truran, Tim Brailsford, Helen Ashman. A Hybrid Technique for English-Chinese Cross Language Information Retrieval, ACM Transactions on Asian Language Information Processing, 7, 2, June 2008.

[30] Ari Pirkola, Turid Hedlund, Heikki Keskustalo, Kalervo Jarvelin. Dictionary-Based Cross-Language Information Retrieval: Problems, Methods and Research Findings, Information Retrieval, 4, 209-230, 2001.