International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-10 Issue-1, November 2020

Cross Language Information Retrieval (CLIR): A Survey of Approaches for Exploring Web Across Languages

Suhas D. Pachpande, Parag U. Bhalchandra

Abstract: ***In the era of globalization, internet being accessible and affordable has gained huge popularity and is widely being used almost everywhere by Government, private organizations, companies, banks, etc. as well as by individuals. It has empowered its users to contribute to the creation of information on web enabling them to use their native languages which consequently has drastically increased the volume of web-accessible documents available in languages other than English. This exponential growth of information on the internet has also induced several challenges before the information retrieval systems. Most of the present monolingual information retrieval systems can retrieve documents in the language of query only, missing the information in other languages that may be more relevant to the user. The need of information retrieval systems to become multilingual has given rise to the research in Cross Language Information Retrieval (CLIR) which can cross the language barriers and retrieve more relevant results from documents in different languages. This article is a review of motivation, issues, work and challenges related to various CLIR approaches. Starting with the most fundamental approaches of translation, it is attempted to study and present a review of more advanced approaches for enhancing the retrieval results in CLIR proposed by various researchers working in this domain.***

Keywords: Cross Language Information Retrieval, Dictionary-Based Translation, Corpus-Based Translation, Machine Translation, lexical ambiguity, bilingual dictionary, term-matching, term frequency, document ranking.

I. INTRODUCTION

Globalization has brought the world together reducing significance of geographical borders for trade as well as information exchange. Internet technologies being more affordable without time and space constraint and easily accessible have enabled the world population to use web as their social and collaboration platform empowering every web user to not only be a web information consumer but also to contribute to creation of information on web. This exponential growth of information on the internet has induced several challenges before the information retrieval systems. "The goal of an information retrieval system is to locate relevant documents in response to a user's query. Documents are typically retrieved as a ranked list, where the ranking is based on estimations of relevance"[1].

Revised Manuscript Received on November 30, 2020.
***** Correspondence Author
Suhas D. Pachpande*, Department of Computer Science, Sant Gadge Baba Amravati University, Amravati (MS), India. Email: suhasdp@gmail.com
Parag U. Bhalchandra, School of Computational Sciences, Swami Ramanand Teerth Marathwada University, Nanded (MS), India. Email: srtmun.parag@gmail.com

Most of the present information retrieval systems are monolingual wherein language of the query and retrieved documents are same. Internet being easily accessible and affordable has been very popular over the last few years and hence most of the government departments, companies, educational institutions and almost all organizations have started using web as their primary storage and communication medium. This information is obviously being transacted in various different languages and is stored in Web documents using multiple languages. As a result, the volume of web-accessible documents available in languages other than English has grown drastically. There might be cases where more precise information relevant to user's request is available in a language other than the language of query. The user may also expect the information to be retrieved in a language in which the user is more comfortable. To facilitate information exchange in this scenario, the information retrieval systems need to be multilingual or cross-lingual. Advances in network architecture have strengthened the infrastructure for information exchange across geographic barriers but are still unable to address the challenges for crossing the language barriers.The monolingual information retrieval engines usually fail to present this information to the user which is against the very basic essence of the ubiquitous world wide web, making information available to user. Users have to manually translate queries which is very inefficient considering time required for translation and constraints due to user's knowledge of unfamiliar languages as well as creating possibilities of retrieving irrelevant information due to incorrect translation.This explosive growth of Internet and diversity of available information sources in several languages has fostered the need for multilingual information retrieval techniques that can cross the language boundaries and has inspired the researchers from Information Retrieval (IR) community to design innovative methodologies for information retrieval across different languages. The Cross Language Information Retrieval abbreviated as CLIR which is a sub domain of Information Retrieval can overcome the language barriers and help retrieving documents in languages that are different from the language of query and offer the most relevant data to the user[2]. Integrating some CLIR tools with traditional search engines may enable them matching terms having same meaning in different languages, presenting this otherwise unexplored data to the user.

"The distinguishing hallmark of a cross language information retrieval (CLIR) engine is the linguistic disparity between the queries which are submitted and the documents which are retrieved. To resolve this disparity, all CLIR engines are required to incorporate some facility for language translation, an obvious requirement if query representations and document representations are to be meaningfully compared"[3].

The most effective and commonly used approaches to Cross Language Information Retrieval (CLIR) include translation for either queries or documents being searched and makes "translation" the most crucial phase in CLIR.

"The two most straightforward techniques for cross-language retrieval are: (1) Automatically translate all documents in the collection into the source language and then apply monolingual retrieval on the translated document collection; and (2) Automatically translate the user-posed query into the target language and then apply monolingual retrieval with the translated query on the original document collection in the target language"[4].

Both approaches to translation have certain limitations. Queries being shorter present limited contexts adding uncertainty and ambiguity to query translation while it may be computationally expensive to translate large document collections. Document translation being computationally expensive, the query translation-based approach which is more flexible and effective has been adopted by most of the approaches to CLIR.

The primary sources of translation knowledge are machine-readable bilingual dictionaries and multilingual corpora. User query can be translated by using methods like Dictionary-Based Translation (DT), Corpus-Based Translation (CT) or Machine Translation (MT).

Using dictionary-based translation, terms from the source language query are replaced by their target language counterparts from the dictionary. Since a single word may have several possible translations, some of these translation alternatives may differ from the meaning intended by the user and may lead to ambiguous translation affecting the retrieval performance of a query.

"Lexical ambiguity is a pervasive problem in natural language processing"[5]. Although very little quantitative information is available about its impact on information retrieval systems, this remains the main hurdle in the entire CLIR process.

In another approach based on parallel corpora, translation knowledge is derived from multilingual text collections using various statistical methods[6]. This approach maps source language documents to target language documents forming a parallel corpus if mapped documents are exact translations of each other and forming comparable corpora if document pairs are not exact translations of each other but share some similarities.

The abundant linguistic resources have contributed to rapid growth for research in the Cross-Language Information Retrieval (CLIR) domain in recent years. Several initiatives like TIDES program supported by DARPA have boosted the research interest in CLIR with a primary objective to make information more comprehensible and readily available.

Research on CLIR has been carried out through various initiatives like TREC, CLEF, NTCIR, etc. focusing on a set of languages other than English such as French, Italian, German, Spanish, Chinese, Arabic, Swedish, Dutch, Finnish, and Russian, etc.

II. RELATED RESEARCH IN CLIR

The main objective of research in information retrieval is not only the model which specifies the way documents and queries are represented and compared to determine relevance estimates; but also focused on finding representations and methods to discriminate between relevant and nonrelevant documents.

Experiments in CLIR have been initiated as early in 1970s by researchers like Salton [1973] working on different models with improved results of retrieval[7].

Indexing the objects to be matched against query terms can retrieve more relevant information and improve the efficiency of information retrieval process. A retrieval model proposed by *Narasimha Raju, Bhadri Raju, Satyanarayana*ranks retrieved documents according to similarity measures representing relevance of documents with user query[8].

Cross-language information retrieval requires matching of term queries with documents in other languages but having similar contextual meaning.Considering the language barrier, direct matching of terms is not possible and hence require translation for expressing the terms from query and documents in the same language; and then employing direct term matching techniques.*Wang and Oard* have introduced a general framework for a meaning-matching model using translation probabilities that integrated knowledge about translation and synonymy [9].

Resolving translation ambiguity of queries is one of the important aspects of cross-language information retrieval. Statistical translation models and relevance language models rely on building parallel bilingual corpora for association and disambiguation of words in a query which is a time consuming and computationally expensive process. The advent of powerful machine-readable dictionaries boosted the research in dictionary-based approaches. In a simplest approach, all translations of each word from query are considered to resolve the problem of ambiguity. This helps retrieving documents relevant to all of these possible translations, although these results may include some less relevant documents due to other meanings of words [10]. Selecting the correct translation of query among all possible translations provided by bilingual dictionary is referred to as the problem of *translation selection*.The word-based model for translation selection uses words as unit for translation which is most simple but also leads to several translation ambiguities. Hence research efforts are focused on choosing phrases that are more specific and reduce the translation ambiguities [11]. A study has been carried out by *Ballesteros and Croft*to determine the role of phrases in query expansion using local context analysis and local feedback with an attempt to reduce errors in automatic dictionary translation [12].

Finding the concepts represented by set of lexical items sharing similar meaning is a complex and error prone task. The framework proposed by *Davidov and Rappoport*, whichutilizes multilingual information from the web to extend concepts represented by a given a set of terms in some language can discover a substantially large number of terms while retaining high precision by disambiguating them using web counts [13].

Retrieval by representing documents using words may lead to irrelevant results due to ambiguity. Representing documents by word senses may improve retrieval performance as they represent semantics of the text creating a basis for lexical semantic relationship useful in construction of thesauri. *Krovetz and Croft* have described experiments on test collections to discover the degree of lexical ambiguity and utility of word senses for discriminating between relevant and nonrelevant documents [14].

Ideally dictionaries should contain words that may occur in documents being searched but actually are static in nature, updated occasionally and may not maintain the vocabulary in live corpora. This gives rise to the out-of-vocabulary (OOV) problem that can have severe impact on the retrieval results in CLIR systems. A hybrid technique for dictionary-based query translation which follows a graph-based model for resolution of candidate term ambiguity with a pattern-based method for the translation of out-of-vocabulary (OOV) terms is proposed by *Zhou, Truran, Brailsford and Ashman*. The experimental results of this hybrid technique performed on NTCIR test collections have shown a substantial increase in retrieval effectiveness over various baseline systems incorporating machine and dictionary-based translation [15].

Although most of the CLIR techniques incorporate token-to-token mappings from bilingual dictionaries, some statistical translation models use multi-term phrases, term dependencies, and contextual constraints yielding better results. A promising approach has been introduced by *Ture, Lin and Oard*, which combines representational advantage of probabilistic structured queries with the richness of the internal representation of a translation model with the help of term translation probabilities generated by alternative query translation [16].

Approaches like measuring the coherence of translated word to the entire query helps to resolve translation ambiguity in dictionary-based CLIR wherein the coherence score of a translated word is computed using word co-occurrence statistics. The algorithm designed by *Pourmahmoud and Shamsfard*selects best translation according to a combination of its coherence with other translated words and word translation probabilities from dictionary [17].

Relevance feedback (RF) is an effective technique in which queries are reformulated before and/or after translation for improving retrieval effectiveness in cross-language information retrieval (CLIR).By selecting better query terms, it can enhance query translation by adjusting translation probabilities thereby resolving some out-of-vocabulary terms. A Relevance Feedback method called translation enhancement (TE) is proposed by *Daqing He and Dan Wu*, "which uses extracted translation relationships from relevant documents to revise the translation probabilities of query terms and to identify extra translation alternatives if available so that the translated queries are more tuned to the current search"[18].

Several online machine translation (MT) systems have empowered the cross-language information retrieval (CLIR) systems for effective query translation. Set of experiments conducted by *Dan Wu, Daqing He, Heng Ji and Grishman*using Google Translate have shown that machine translation (MT) is an excellent tool for query translation which when coupled with relevance feedback can achieve significant improvements over the monolingual baseline. It is further found that MT based query translation is effective for both short as well as long queries [19].

Unavailability of test corpora and computational cost for building corpora, limits the effectiveness of parallel corpus approach to CLIR,inviting research for tapping potential of other knowledge based approaches.*Sadat, Maeda, Yoshikawa and Uemura*have worked on query translation and disambiguation to improve the effectiveness of information retrieval by combining statistical disambiguation method both before and after translation. Evaluation of the disambiguation method applied to French-English Information Retrieval using TREC data collection has proved the effectiveness of this method [20].

In a Web directory based CLIR method proposed by *Kimura, Maeda, Yoshikawa and Uemura*, feature terms are extracted from web documents for each category in the source and target languages. By comparing a particular category in a language with categories across different languages, one or more corresponding categories in another language are determined based on the similarity index. Results obtained by testing various methods including employing corpus statistics for the translation of terms and disambiguation show that these methods are heavily affected by the domain of training corpus significantly reducing the retrieval effectiveness for other domains [21].

A query disambiguation method independent of a particular domain using Web directory as the corpus is proposed by *Kimura, Maeda, Hatano, Miyazaki and Uemura*. It is based on estimating for domains of the query using hierarchic structures of Web.Web directories written in many different languages are utilized as "multilingual corpus for disambiguating translation of the query and for estimating the domain of search results using hierarchic structures of Web directories"[22]. Experimental evaluations have shown improvement in retrieval accuracy effectively restricting the target fields of query using lower level merged categories and thereby obtaining most suitable translation of the query. This method is effective in especially when the document collection includes a wide range of domains like the Web.

Corpus-based disambiguation methods are used to translate terms and their disambiguation but are significantly affected by the domain of the training corpus as CLIR methods need to be independent of specific domains. Variations in web page layouts and writing styles make it difficult to extract term translations from bilingual web pages that may consist of rich pool of term translation knowledge.

Based on the observation that translation pairs on the same web page follow similar patterns, a new extraction model has been proposed by *Lei Shi*, which adaptively learn extraction patterns and exploit them to facilitate term translation mining from bilingual web pages. Experiments on TREC Collections reflect that this adaptive web mining model improves query translation providing better extraction coverage leading to significantly effective retrieval with higher accuracy [23].

As the effectiveness of translation models usually suffers due to the lack of large parallel corpora, *Jiang Chen and Jian-Yun Nie*have described a parallel text mining system that searches for parallel texts between Chinese and English automatically on the Web and this parallel corpus is then used to train a probabilistic translation model for translating queries.The feasibility of statistical translation model has been investigated and it is found that the generated evaluation lexicons providedresults with better precision using this relatively noisy Chinese-English parallel corpus constructed from the Web [24].

Results obtained by studies on translation of terms using corpus statistics alongwith disambiguation reveal that domain of training corpus affects the corpus-based disambiguation methods, significantly reducing the retrieval effectiveness for other domains. Hence a CLIR method employing a Web directory provided in multiple language versions is proposed by *Kimura, Maeda, Yoshikawa and Uemura*, which extracts feature terms from web documents for each category in both the source and target languages and compare the similarities between categories across languages.Ambiguities generated using dictionary translation can be reduced with these category pairs by narrowing the categories to be retrieved from the target language. The proposed method which is independent of a particular domain uses a Web directory for CLIR and is effective since it covers wide range of domains such as the Web in its document collection [25].

*Tholpadi, Bhattacharyya and Shevade*have explored using auxiliary language corpora for translation induction and semantic relatedness measurement in the cross-lingual domain and by creating two new human-annotated datasets have demonstrated significant gains in the performance for 21 language pairs when using auxiliary languages for the CC-CLSR task [26].

III. APPROACHES TO CLIR

*Sadat et al.*have focused on a combined approach for query translation and disambiguation using methods that do not rely on scarce resources such as parallel corpora but use Bilingual Machine Readable Dictionaries (MRDs) as an alternative. Using ranking and selection of source query terms, this combined statistical disambiguation approach applied before and after translation, can select best target translations and significantly reduce errors generated by simple dictionary translation [20].

For English-Hindi based CLIR system, *Katta and Arora* have applied Naïve Bayes and particle swarm optimization algorithms to improve ranking and searching aspects of a CLIR system by matching terms contained in documents to the query terms in same sequence as present in the search query. A basic English-Hindi CLIR system which uses English to Hindi translator, a synonym generator and a query suggestion part; is improvised by incorporating PSO and Naïve Bayes algorithms to the system for effective searching and ranking using n-gram matching.A bilingual English-Hindi translator is used in this approach alongwith Hindi query extension and synonym generation which retrieved more relevant results in an English-Hindi based CLIR [27]. An approach involving pre and post query expansion is proposed to improve the performance of English-Hindi CLIR system using Local Expansion using initial query, definition based pre query expansion and keyword ranking. Experiments of the proposed approach performed on FIRE 2010 (Forum of Information Retrieval Evaluation) datasets have shown significant improvements in performance of English-Hindi CLIR system in terms of average precision[28]. A hybrid approach for Persian-English CLIR is proposed which exploits “a combination of phrase reorganization, pattern based phrase translation and query expansion before and after translation along with a probabilistic algorithm to choose the best translation of words and phrases”, finally ranking documents according to statistical language model[17]. A method for creating a comparable text corpus has been presented by *Talvensaari et al.*wherein keys with best resolution power are extracted using the relative average term frequency (RATF) value from the source document collection. These keys are then translated into the language of target collection using a dictionary-based query translation method. Executing these translated queries against target document collection, formed comparable collection containing alignment pairs for the retrieved documents that matched a similarity score criteria, which further was used as a similarity thesaurus to translate queries in addition to a dictionary-based translator. This approach aligns two document collections in different languages as a top-N-ranking document hierarchy.The effectiveness of the approach is tested with several combined CLIR approaches based on comparable corpora, dictionary-based query translation, and pseudorelevance feedback. The results have shown that this combined approach outperformed translation schemes using either dictionary-based translation or corpus translation only [6].

There are three possible approaches to translation including use of machine translation (MT) system, bilingual dictionary, and a statistical/probabilistic model based on parallel texts.MT systems primarily focus on producing syntactically correct sentences usually selecting one of several possible translations of a word which may end up selecting a wrong equivalent of the word in target language leading to retrieval of irrelevant results.

Every bilingual dictionary usually has many translations of a word reflecting different meanings that may cause ambiguities and may add substantial noise to the retrieved results.

Another approach based on the corpus of parallel texts establishes translation correspondences between group of words or sentences with their target equivalent which enables determining the most probable translation of a word.

This approach does not require any bilingual dictionary or a MT system and can be made specific to a domain thereby reducing noises and improving retrieval results.*Nie, Simard, Isabelle and Durand* have investigated a flexible probabilistic translation model constructed by automatically gathering parallel texts from the Web for English-French languages in order to construct a reasonable training corpus. The results obtained after evaluation of the model applied on 5000 documents are close to that of using an MT system and can be used for CLIR where appropriate parallel corpus is available. Although the probabilistic translation approach using parallel corpora may eliminate the need of using a MT system, it usually suffers from a major obstacle of unavailability of parallel texts [29].

Another approach to resolve translation ambiguity is by computing the coherence of a translation word to the entire query which is computed using word co-occurrence statistics. Liu, Jin and Chai have proposed a statistical model named Maximum Coherence Model, which estimates the translation probabilities of query words that are consistent with the word co-occurrence statistics. This model estimates the translation probabilities of query words which resolves the translation uncertainty problem and drops the translation independence assumption by estimating translation probabilities for all query words simultaneously. Evaluation of the model with TREC datasets have shown substantial improvement over other selection-based approaches using word co-occurrence statistics for sense disambiguation [10].

A CLIR method is proposed which employs two or more language versions of a Web directory, one of which is the query language and others are the target languages. Category correspondences between languages are estimated in advance by extracting feature terms for each category in both source and target languages. Category pairs are formed across languages by comparing similarities among them; which are helpful in resolving ambiguities of simple dictionary translation [21]. A simple cost-effective corpus-based dictionary construction approach using a word-extraction algorithm which utilizes local contextual information can outperform static dictionary and the bigram indexing approach mainly because the constructed dictionary size is relatively smaller than that of a static dictionary.*Jin and Wong*constructed Chinese dictionaries from different Chinese corpora and applied the words from these dictionaries for indexing in information retrieval. Results obtained for three different Chinese corpora have shown effective retrieval using indexes based on the constructed dictionary. During translation in cross lingual information retrieval, inference is primarily based on word relations and hence word-based indexing might be a most suitable approach [30].

An approach involving the Hidden Markov Models is proposed by *Jinxi Xu and Weischedel*, based on the probability of generation of a query in one language from a document in another language. This integrated probabilistic model requires only a bilingual dictionary as a resource and combines probability model of term translation and retrieval. Results of experiments using English queries to access Chinese documents (TREC-5 and TREC-6) and Spanish documents (TREC-4) show effectiveness of the model using bilingual dictionary by reducing the performance degradation caused due to translation ambiguity. The method uses Hidden Markov Model (HMM) instead of using tf-idf to estimate the probability of a document relevance against the given query, integrating two simple estimates of term translation probability into the monolingual HMM [31].

Several experiments were performed on NTCIR-4 test collections by *Jacques Savoy*with an objective to review the retrieval effectiveness of various vector-space and probabilistic models used for monolingual searches and to determine effectiveness of various automated and freely available tools to translate English queries into some of the Asian languages to retrieve pertinent documents in these languages.The evaluation results demonstrated that the "Lnu-ltc" vector space model and the Okapi probabilistic IR model show better retrieval performance levels when indexing Asian languages based on bigrams and the Blind-query expansion approach is proven to be worthwhile especially for processing short queries [32].

Distributed word vector representations are being popular in Natural Language Processing (NLP) tasks and are used to identify similar contextual words. Effectiveness of word vectors in Hindi-English CLIR has been analyzed by*Sharma and Mittal*using SkipGram Model (SGM) to learn bi-lingual word vectors from sentence aligned comparable corpus, aligning the source and target language words from the corpus using IBM model and finally selecting best target language translation with the help of top-n word alignments and word vectors [33].

IV. CONCLUSION

The exponential growth of information on web in multiple languages and the availability of abundant linguistic resources have contributed to rapid growth for research in the Cross-Language Information Retrieval (CLIR). Beginning with manual translation of queries researchers have presented various approaches for automatic translation of queries and documents including Dictionary-Based Translation (DT), Corpus-Based Translation (CT) and Machine Translation (MT). Each of these approaches have certain limitations and challenges during implementation, inviting further research in this domain. Ambiguity is one of the major challenges faced in dictionary-based translation approach, building parallel corpus is a bottleneck in the successful implementation of corpus-based translation while availability of a powerful machine translation tool is a hurdle in success of machine translation approach. Accepting the challenge, several researchers have developed and evaluated numerous models and approaches to overcome these challenges and have presented very promising results that enhance the retrieval outcomes. In this article we have attempted to review some of these approaches like using indexing, phrases, context analysis, relevance feedback, term dependency, coherence score and some statistical approaches. Some approaches using tools like Google Translate, web directories and repositories, bilingual machine-readable dictionaries, etc. are also studied.

The results obtained by the researchers by evaluating approaches involving use of Naïve Bayes and particle swarm optimization algorithms, relative average term frequency, Probabilistic Translation Model, Maximum Coherence Model, category correspondences, Hidden Markov Model, etc. demonstrate improvement in retrieval.

The effectiveness of vector representations of words and phrases using some of the vector space models and probabilistic models for information retrieval is very promising and may be explored further. Blending some of the approaches presented earlier with Vector Space Model, which is one of the most famous and effective models for monolingual information retrieval would certainly enhance accuracy of retrieval results in Cross Language Information Retrieval.

REFERENCES

1. Robert Krovetzand W. Bruce Croft, *"Lexical Ambiguity and Information Retrieval"*, ACM TransactIons on Information Systems, Vol. 10, No 2, April 1992.
2. Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Chung Tsai, *"Proper Name Translation in Cross-Language Information Retrieval"*, ACL '98/COLING '98: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17thInternational Conference on Computational Linguistics - Volume 1 August 1998 Pages 232–236.
3. 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, Vol. 7, No. 2, Article 5, June 2008.
4. Christof Monz, Bonnie J. Dorr, *"Iterative Translation Disambiguation for Cross-Language Information Retrieval"*, SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval August 2005 Pages 520–527.
5. Robert Krovetz, W. Bruce Croft, *"Lexical Ambiguity and Information Retrieval"*, ACM Transactions on Information Systems, Vol 10, No 2, April 1992, Pages 115-141.
6. TuomasTalvensaari, JormaLaurikkala, KalervoJarvelin, MarttiJuhola, HeikkiKeskustalo, *"Creating and Exploiting a Comparable Corpus in Cross-Language Information Retrieval"*, ACM Transactions on Information Systems, Vol. 25, No. 1, Article 4, February 2007.
7. Jiang Chen and Jian-Yun Nie, *"Automatic construction of parallel English-Chinese corpus for cross-language information retrieval"*, ANLC'00: Proceedings of the sixth conference on Applied natural language processing, April 2000, Pages 21–28.
8. B.N.V Narasimha Raju, Dr. M S V S Bhadri Raju, Dr. K V V Satyanarayana, *"Translation Approaches in Cross Language Information Retrieval"*, IEEE Proceedings: International Conference on Information Technology: Coding and Computing, 2002.
9. Jianqiang Wang and Douglas W. Oard, *"Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval"*, SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, August 2006, Pages 202–209.
10. Yi Liu, Rong Jin, Joyce Y. Chai, *"A Maximum Coherence Model for Dictionary based Cross language Information Retrieval"*, SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, August 2005, Pages 536–543.
11. Jianfeng Gao, Jian-Yun Nie, *"A Study of Statistical Models for Query Translation: Finding a Good Unit of Translation"*, SIGIR '06:0 Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, August 2006, Pages 194–201.
12. Lisa Ballesteros and W. Bruce Croft, *"Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval"*, ACM SIGIR Forum, July 1997, Pages 84-91.
13. Dmitry Davidov, Ari Rappoport, *"Enhancement of Lexical Concepts Using Cross-lingual Web Mining"*, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6-7 August 2009, pages 852–861.
14. Robert Krovetz, W. Bruce Croft, *"Lexical Ambiguity and Information Retrieval"*, ACM Transactions on Information Systems, Vol 10, No 2, April 1992, Pages 115-141.
15. 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, Vol. 7, No. 2, Article 5, June 2008.
16. Ferhan Ture, Jimmy Lin, Douglas W. Oard, *"Looking Inside the Box: Context-Sensitive Translation for Cross-Language Information Retrieval"*, SIGIR '12: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, August 2012, Pages 1105–1106.
17. SolmazPourmahmoud, MehrnoushShamsfard, *"Semantic Cross-lingual Information Retrieval"*, IEEE 23rd International Symposium on Computer and Information Sciences, Nov 2008.
18. Daqing He, Dan Wu, *"Translation Enhancement: A New Relevance Feedback Method for Cross-Language Information Retrieval"*, CIKM '08: Proceedings of the 17th ACM conference on Information and knowledge management, October 2008, Pages 729–738.
19. Dan Wu, Daqing He, Heng Ji, Ralph Grishman, *"A Study of Using an Out-Of-Box Commercial MT System for Query Translation in CLIR"*, iNEWS '08: Proceedings of the 2nd ACM workshop on Improving non english web searching, October 2008, Pages 71–76.
20. F. Sadat, A. Maeda, M. Yoshikawa, S. Uemura, *"A combined statistical query term disambiguation in cross-language information retrieval"*, IEEE Proceedings:13th International Workshop on Database and Expert Systems Applications, 2002, Pages 251-255.
21. F. Kimura, A. Maeda, M. Yoshikawa and S. Uemura, *"Cross-language information retrieval using Web directories"*, 2003 IEEE Pacific Rim Conference on Communications Computers and Signal Processing (PACRIM 2003) (Cat. No.03CH37490), vol.2, 2003, Pages 911-914.
22. F. Kimura, A. Maeda, K. Hatano, J. Miyazaki, S. Uemura, *"Cross-Language Information Retrieval by Domain Restriction Using Web Directory Structure"*, IEEE Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), 2008, Pages 135-135.
23. Lei Shi, *"Adaptive Web Mining of Bilingual Lexicons for Cross Language Information Retrieval"*, CIKM '09: Proceedings of the 18th ACM conference on Information and knowledge management, November 2009, Pages 1561–1564.
24. Jiang Chen, Jian-Yun Nie, *"Automatic construction of parallel English-Chinese corpus for cross-language information retrieval"*, ANLC '00: Proceedings of the sixth conference on Applied natural language processing, April 2000, Pages 21–28.
25. F. Kimura, A. Maeda, M. Yoshikawa, S. Uemura, *"Cross-Language Information Retrieval Based on Category Matching Between Language Versions of a Web Directory"*, AsianIR '03: Proceedings of the sixth international workshop on Information retrieval with Asian languages, Volume 11 July 2003, Pages 153–160.
26. Goutham Tholpadi, Chiranjib Bhattacharyya, Shirish Shevade, *"Corpus-Based Translation Induction in Indian Languages Using Auxiliary Language Corpora from Wikipedia"*, ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 16, No. 3, Article 20, March 2017.
27. Eva Katta, Anuja Arora, *"An Improved approach to English-Hindi based Cross Language Information Retrieval System"*, IC3 '15: Proceedings of the 2015 Eighth International Conference on Contemporary Computing (IC3), August 2015, Pages 354–359.
28. S. Varshney and J. Bajpai, *"Improving Retrieval performance of English-Hindi based Cross-Language Information Retrieval"*, 2013 IEEE International Conference in MOOC, Innovation and Technology in Education (MITE), Jaipur, 2013, Pages 300-305.
29. Jian-Yun Nie, Michel Simard, Pierre Isabelle, Richard Durand, *"Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web"*, SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, August 1999, Pages 74–81.
30. HonglanJin, Kam-Fai Wong, *"A Chinese Dictionary Construction Algorithm for Information Retrieval"*, ACM Transactions on Asian Language Information Processing, Vol. 1, No. 4, December 2002, Pages 281-296.
31. Jinxi Xu, Ralph Weischedel, *"Cross-lingual information retrieval using hidden Markov models"*, EMNLP '00: Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13,

October 2000, Pages 95–103.
32. Jacques Savoy, *"Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages"*, ACM Transactions on Asian Language Information Processing, Vol. 4, No. 2, June 2005. Pages 163-189.
33. Vijay Kumar Sharma, Namita Mittal, *"Exploring Bilingual Word Vectors for Hindi-English Cross Language Information Retrieval"*, ICIA-16: Proceedings of the International Conference on Informatics and Analytics, August 2016 Article No.: 28, Pages 1–4.

AUTHORS PROFILE

Suhas D. Pachpandeis Assistant Professor in Department of Computer Science of Sant Gadge Baba Amravati University, Amravati (MS) India. He has completed his MSc (Computer Science) from North Maharashtra University, Jalgaon. His research interest includesWeb mining, Information Retrieval.

Dr. Parag Bhalchandrais Assistant Professor in School of Computational Sciences of Swami Ramanand TeerthMarathwada University, Nanded (MS) India.He has 15+ years teaching experience and has 30+ research publications to his account. His research areas include Analytics, Compilers and ICT.