RESEARCH ARTICLE OPEN ACCESS

Marathi-English CLIR using detailed user query and unsupervised corpus-based WSD

Savita C. Mayanale*, Ms. S. S. Pawar**
* (M.E. Student, Dept. of Computer Engg., D. Y. Patil College of Engg., Akurdi, Savitribai Phule Pune University, India.)
** (Asst. Prof., Dept. of Computer Engg., D. Y. Patil College of Engg., Akurdi, Savitribai Phule Pune University, India.)

ABSTRACT
With rapid growth of multilingual information on the Internet, Cross Language Information Retrieval (CLIR) is becoming need of the day. It helps user to query in their native language and retrieve information in any language. But the performance of CLIR is poor as compared to monolingual retrieval due to lexical ambiguity, mismatching of query terms and out-of-vocabulary words. In this paper, we have proposed an algorithm for improving the performance of Marathi-English CLIR system. The system first finds possible translations of input query in target language, disambiguates them and then gives English queries to search engine for relevant document retrieval. The disambiguation is based on unsupervised corpus-based method which uses English dictionary as additional resource. The experiment is performed on FIRE 2011 (Forum of Information Retrieval Evaluation) dataset using "Title" and "Description" fields as inputs. The experimental results show that proposed approach gives better performance of Marathi-English CLIR system with good precision level.
Keywords **-** Morphological Analyser, CLIR, Dictionary-based Approach, word sense disambiguation.

I. INTRODUCTION

Information Retrieval (IR) refers to a process that users can find their required information or knowledge from corpus including different kinds of information. Traditional IR system is monolingual system. However, with the rapid growing amount of information available to us, the situations that a user needs to use a retrieval system to perform querying a multilingual document collection are becoming increasingly emerging and common. This tendency causes the difficulty of information acquisition. Meanwhile, language barriers become a serious problem. As a result CLIR has received more research attention and is increasingly being used to retrieve information on the Internet.

India has multiple languages which are written in different scripts. Most of the people in India speak in their regional language whereas vast amount of information on the Internet is available in English. However, users who do not use English as first language are also significantly high in number. Such non English users find it difficult to express the query in appropriate English language. Proficiency in English language is becoming a kind of barrier in finding rich source of information available on World Wide Web. CLIR helps in bridging this gap. CLIR systems allow users to query in their native language and get relevant documents in other languages.

Translations in CLIR can be achieved by two ways: Query translation or Document translation [16]. In Query Translation, the given query will be translated from native language to target language and then search operation is performed to get the relevant documents. In document translation, all the documents from corpus are translated into native language. It allows the user to ask query in native language and then the searching will take place to obtain the relevant documents in native language. Among the two, the query translation is easier compared with document translation because of the size of translation [8]. But, the drawback with query translation is that the given query normally will be short and hence ambiguity problem may arise.

These translation methods can be further classified into three classes according to what resources are used to cross the language boundary as Machine Translation (MT), dictionary based and parallel corpora based. In [1], CLIR experiments were conducted using Google Translator for translating queries to study the effectiveness of CLIR using MT for query translation. In [5] and [13], CLIR systems were implemented using bilingual dictionaries. Parallel or comparable corpora are useful resources to extract beneficial information for CLIR [14].

The performance of CLIR system mainly depends on accurate translation of users input query. But, most of the times user's query contains ambiguous terms which results in retrieval of irrelevant documents also along-with relevant

documents. This reduces the overall precision. In order to improve the retrieval efficiency; Word Sense Disambiguation (WSD) is used. WSD approaches are classified in two major categories depending on the corpora they use. Supervised approaches need sense tagged corpora for training the model whereas unsupervised approaches work on untagged corpora. There are some semi-supervised approaches which work with small annotated seed data. Another way to classify WSD approaches is based on the use of knowledge resources. Approaches using knowledge resources like wordnet, ontologies, thesauri are termed as knowledge based approaches, while there are some approaches which work without any of these resources, which are termed as knowledge-lean or corpus based approaches.

II. RELATED WORK

The Kazuaki Kishida [2] in 2004 reviewed techniques and methods for enhancing effectiveness of cross-language information retrieval. They have covered research issues such as (i) matching strategies and translation techniques, (ii) methods for solving the problem of translation ambiguity, (iii) formal models for CLIR such as application of the language model, (iv) the pivot language approach, (v) methods for searching multilingual document collection, (vi) techniques for combining multiple language resources, etc. Zhang Tao and Yue-Jie Zhang in 2007 and 2008 implemented two systems based on the work in CLIR evaluation task at the 9th Text Retrieval Conference (TREC-9): English-Chinese [3] and Chinese-English [4]. The machine readable dictionaries are utilized as the important knowledge source to acquire correct translations for English-Chinese and Chinese-English.

Chinnakotla Manoj Kumar et al. [5] in 2008 developed Marathi-English CLIR system which uses bi-lingual dictionaries and query translation approach. Simple rule based approach is used to transliterate the query words which are not found in the dictionary. The resulting multiple translation/transliteration choices for each query word are disambiguated using an iterative page-rank style algorithm to produce the final translated query. Almeida Ashish and Bhattacharyya P. [6] in 2008 studied the effects of lexical analysis on Marathi monolingual search over the news domain corpus of FIRE-2008 and observed the effect of processes such as lemmatization, inclusion of suffixes in indexing and stop words elimination on the retrieval performance.

Sethuramalingam S et al. [7] in 2009 conducted CLIR experiments between three languages which uses writing systems (scripts) of Brahmi-origin, namely Hindi, Bengali and Marathi. The similarity of the writing systems used for these Indian languages can be used effectively to improve CLIR and also to overcome the problems of textual variations, textual errors, and even the lack of linguistic resources like stemmers to an extent. The results were significantly improved for all the six language pairs using a method for fuzzy text search based on Surface Similarity. Nasharuddin et al. [8] in 2010 reviewed some recent researches focusing on topics in cross lingual information retrieval and its role in research directions which include new models and paradigms in the wide area of information retrieval. In [9], various approaches for CLIR were explained in detail.

Kumar Sourabh [10] in 2013 conducted an extensive literature review on various developments in India in machine translation system and Cross Language IR. This survey paper covers ongoing developments in CLIR and MT with respect to Status of the projects, Current projects, Participants, Existing projects, Government efforts, Funding and financial aids, Twelth Five Year Plan (2012-2017) Projections, Eleventh Five Year Plan (2007-2012) activities. Arora P. et al. [11] in 2013 described details of DCU's participation in the Cross Language !ndian News Story Search task (CL!NSS) at FIRE 2013. This task is an edition of the PAN@FIRE task which focuses on addressing news story linking between English and Indian languages.

H. B. Patil et al. [12] in 2014 demonstrated a rule-based Part-of-Speech tagger for Marathi Language. The hand constructed rules that are learned from corpus and some manual addition after studying the grammar of Marathi language are added and that are used for developing the tagger. Disambiguation is done by analyzing the linguistic feature of the word, its following word, its preceding word etc. S. Varshney and J. Bajpai [13] in 2014 proposed an algorithm for improving the performance of English-Hindi CLIR system. They used all possible combination of Hindi translated query using transliteration of English query terms and choosing the best query among them for retrieval of documents. The experiment is performed on FIRE 2010 (Forum of Information Retrieval Evaluation) datasets.

Pratibha Bajpai et al. [14] in 2014 analyzed the various researchers work in the area of Indian language CLIR. They had also presented prototype for English to Hindi language CLIR and issues related to the English to Hindi language translation. The authors had tested 30 queries manually using suggested prototype and found that the precision level is quite good. Alessio Bosca et al. [15] in 2014 presented a CLIR system exploiting multilingual ontologies for enriching documents' representation with multilingual semantic information during the indexing phase and also for mapping query fragments to concepts during the retrieval phase.

This system has been applied on a domain specific document collection and the contribution of the ontologies to the CLIR system has been evaluated in conjunction with the use of both Google and Microsoft Bing translation services. Results demonstrated the use of domain-specific resources which lead to a significant improvement of CLIR system performance.

Sandhan [17] is a monolingual search system developed under Technology Development for Indian Languages (TDIL) programme in tourism domain for five Indian languages viz. Bengali, Marathi, Hindi, Tamil and Telugu. In this system, user has the facility to submit query using the INSCRIPT or phonetic layout. It processes the query based on its language and retrieves results from the respective language. Summary is generated for each retrieved document and this feature helps the user in knowing the basic information about the overall content of the document without opening it.

III. PROPOSED METHODOLOGY

Marathi is one of the widely spoken languages in India especially in the state of Maharashtra. Marathi uses the Devanagari script and draws vocabulary mainly from Sanskrit [5]. The architecture of our Marathi-English CLIR system is shown in Fig. 1. The system is based on query translation approach as it is not feasible to translate documents. The input to the system is Marathi topic and description given in FIRE 2011 topic set. As Marathi is morphologically rich like other Indian languages, there is need to stem the query words to get the Marathi root words. Query is translated into English using bilingual dictionary and transliterated using simple mapping approach from source language characters to the target language characters. Then all possible combinations of translations are disambiguated and given to the search engine.

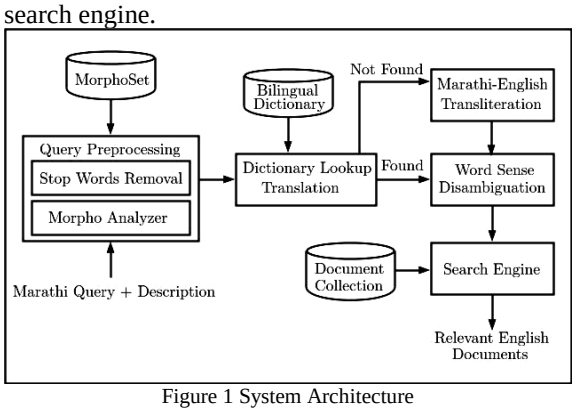

Figure 1 System Architecture

A. Preprocessing of query

The system takes Marathi query as an input. This query is first pre-processed to remove stop words and to get root words. Stop words are commonly used words that a search engine has been programmed to ignore, both when indexing and retrieving documents as the result of a search query such as the, is, at, which, on, etc. Then the query words are stemmed before looking up their entries in the bi-lingual dictionary which is called as Morphological Analysis. A simple lookup approach is used to remove stop words and to get root words.

B. Query Translation

Translation of query from one language to other language is known as query translation. Query translation is a crucial step in CLIR system because all problems come from this step like mismatching of query terms, ambiguities, poor retrieval performance etc. The system uses machine-readable bi-lingual Marathi-English dictionary for query translation. Bilingual dictionaries are specialized in translating text and words from one language to another. It is easy to translate query by replacing words with its equivalent word available in bilingual dictionary.

C. Transliteration

It is a process of converting any text or words from one script to another. Many tools are available for transliteration into English from Devanagari format such as Itrans, IAST, etc. If a translation for input query is not found in dictionary, the word is assumed to be a proper noun and therefore transliterated by the Marathi-English transliteration module. The module is based on simple mapping approach which returns appropriate English character transliteration for each Marathi character.

D. Word Sense Disambiguation

In this paper, we propose a new unsupervised corpus-based WSD algorithm which uses similarity between senses of the target word and a context vector. It also uses English dictionary as additional resource to find synonyms for correctly sensed word so that more relevant documents can be retrieved. The sense of the target word is determined as that for which the similarity between gloss and context vector is greatest. WSD is used to detect the correct sense of the Marathi word in English. This module makes use of the English Barcelona corpus and English dictionary available at [19]. The Barcelona corpus is preprocessed by removing stop words and then it is converted to excel database for easy manipulation of the corpus.

Let us consider a query "गोधा जलद ट्रेन". After preprocessing, translation and transliteration, we get two meanings for word जलद from bilingual dictionary: "fast and cloud". So, two translations can be formed, first "godhra fast train" and second "godhra cloud train". Initially "godhra fast train" is fed to WSD module. Now in WSD module, system will search for each individual word of input translation in excel corpus. If "train" is found in one of the rows, then that row is locked and further search is made for "godhra" and "fast" in the locked

row. Now if we find "fast" in the locked row then "fast" is marked as correct sense. After this, "fast" is searched again in the English-English dictionary to get its synonyms. For the word "fast"; rapid, speedy, etc. will be retrieved as synonyms. Thus by searching correct sensed word in the English dictionary, more relevant documents are retrieved. The same procedure would be carried for "godhra cloud train" but no correct sense is obtained for this translation in excel corpus. Finally three translated queries are submitted to the search engine: godhra fast train, godhra rapid train and godhra speedy train.

E. Monolingual English IR

The possible disambiguated translations of Marathi query are submitted to the English IR engine. Apache Lucene is used for indexing the input documents and searching the queries over target collection. Lucene creates Inverted Index. Normally, we map document => terms in the document. But, Lucene does the reverse. It creates index of terms => list of documents containing the term, which makes it faster to search. Lucene combines Boolean model (BM) of Information Retrieval with Vector Space Model (VSM) of Information Retrieval documents "approved" by BM are scored by VSM [20].

The proposed algorithm is given below that shows the step by step process of Marathi-English CLIR.

Algorithm 1 Marathi-English CLIR by combining translation and transliteration

Input: Query and description in Marathi script

Output: Relevant documents in English with respect to the given query

1: Remove stop words from user query and get the root words in Marathi
2: for each root word **do**
3: Translate into English language using Marathi-English bilingual dictionary
4: if not found in dictionary **then**
5: Transliterate into English language
6: Make all possible translations in English using step 3 and 5
7: Disambiguate all translations for each word
8: Submit disambiguated queries to English search engine
9: Display all relevant English documents generated to the users.

IV. DATASET AND RESOURCES

A. Document Collection

The FIRE (Forum of Information Retrieval Evaluation) 2011 dataset is English document collection consisting of news articles from "The Telegraph, Calcutta Edition" from 2001-2010. For this system we have considered news articles from year 2004-07. Dataset consists of set of user queries in terms of "Title" field, "Description" field and "Narrative" field, set of documents and "qrel" files which gives a list of relevant documents for queries. The system uses Marathi queries to retrieve English documents. For the proposed system, we used "Title" and "Description" field of the queries. Dataset is available online for free at [18].

B. CFILT Resources

The system uses various resources created by Center for Indian Language Technologies (CFILT) such as bilingual Marathi-English dictionary for translation, Morphological analyzer to get the root words in Marathi, Barcelona corpus and English dictionary for word sense disambiguation. All resources are available online at [19].

C. Apache Lucene

The proposed system uses open source Apache Lucene [20] search engine library as English IR engine, i.e. indexing the input documents and searching relevant documents over the target collection.

V. EXPERIMENTAL RESULTS

We have developed two systems: CLIR1 and CLIR2 which are Marathi-English CLIR based on query translation approach. Both systems takes query in Marathi language which is translated into English using bilingual dictionary after pre-processing and the words which are not found in dictionary are transliterated. The final query is given to search engine to retrieve relevant documents in English. The first system developed in this work has no WSD module whereas second system contains WSD. In addition to the input query, second system takes two more inputs: description of a query and the domain in which the system should retrieve results. TABLE I shows statistics of data collection used for implementation. TABLE II shows results for 15 queries which are tested on proposed system along with graphs.

TABLE I Statistics of Data Collection

SN	Metrics	CLIR
1	Query Language	Marathi
2	Document Language	English
3	No. of queries	5
4	No. of Documents	1,20,379
5	Size of Collection	288MB
6	Avg. No. of relevant documents per query	15

TABLE II Precision and Recall Results

SN	Query	CLIR1		CLIR2	
		Recall	Precision	Recall	Precision
Q1	स्वाईन फ्लू लस	1	0.03	1	0.03
Q2	गोध्रा ट्रेन आक्रमण	0.617021	0.145	0.851064	0.2
Q3	मायकल जॅक्सनचा अकाली मृत्यू	1	0.025	1	0.025
Q4	बराक ओबामाचा विजय	0.785714	0.055	1	0.07
Q5	अबू घारिब जेलमधील छळ	0.685714	0.12	0.857143	0.15
Q6	मनोरंजनाच्या जगातील चाचेगिरी	0.90625	0.145	0.9375	0.15
Q7	भारतीय महिला आरक्षण विधेयक	0.666667	0.05	0.866667	0.065
Q8	सोमाली समुद्री चाच्यांचा पूर्णपणे पराभव	1	0.055	1	0.055
Q9	क्लोन केलेल्या मानवी मूलांचा जन्म	1	0.055	1	0.055
Q10	बेकायदेशीर झाडे तोडणे	0.486486	0.18	0.675676	0.25
Q11	भोपाळ वायू दुर्घटना	0.961538	0.25	1	0.26
Q12	बेनजीर भुट्टो यांची हत्या	0.647059	0.055	0.882353	0.075
Q13	भारतातील सायबर गुन्हा	0.961538	0.125	0.961538	0.125
Q14	दिल्लीमधील राष्ट्रकुल खेळ	0.666667	0.06	0.888889	0.08
Q15	बिल गेट्सचे परोपकारी प्रयत्न	1	0.025	1	0.025

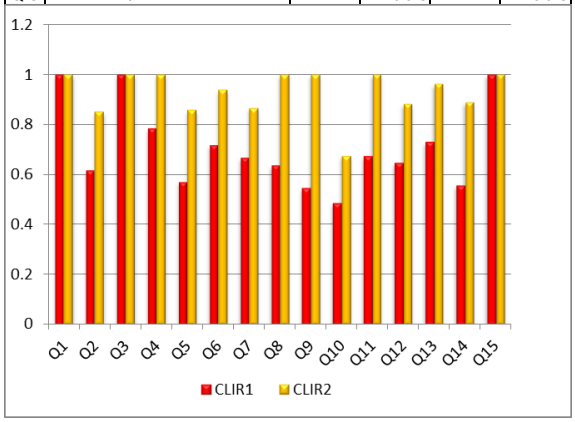

Figure 2 Recall Results

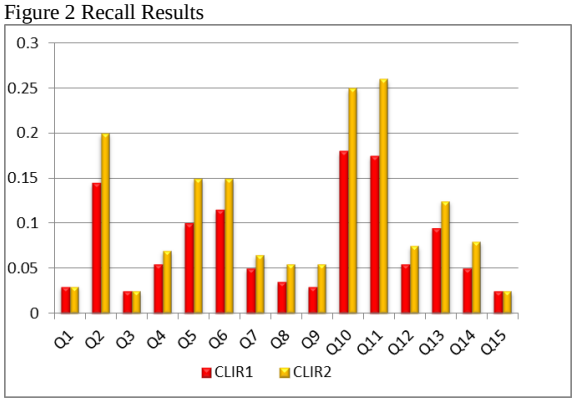

Figure 3 Precision Results

VI. CONCLUSION

Cross language IR is a technique for searching documents in many languages across the world and it can be the baseline for searching not only among two languages but also in multiple languages. In this paper, we have proposed a new approach for word sense disambiguation of Marathi-English CLIR system which makes use of corpus and dictionary in excel format. In addition to query, the user can provide description and domain for the query. The approach of using detailed query performed well and retrieved relevant results with good precision.

In future, we would like to explore alternative search engines and scoring functions instead of standard Lucene scoring function.

Acknowledgements

We would like to thank the publishers as well as researchers for making their resources available and teachers for their guidance. We are thankful to authorities of Savitribai Phule Pune University for their constant guidelines and support. We also thank the college authorities for providing the required infrastructure and support. Lastly, we would like to extend a heartfelt gratitude to friends and family members who have inspired, guided and assisted us in performing this work.

REFERENCES

[1] D. Wu and D. He, "A study of query translation using google machine translation system," in Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on, pp. 1–4, IEEE, 2010.

[2] K. Kishida, "Technical issues of cross-language information retrieval: a review," Information processing & management, vol. 41, no. 3, pp. 433– 455, 2005.

[3] Zhang Yue-Jie and Tao Zhang, "Research on English-Chinese Cross- Language Information Retrieval," International Conference on Machine Learning and Cybernetics. IEEE, 2007.

[4] Zhang Tao and Yue-Jie Zhang, "Research on Chinese-English Cross- Language Information Retrieval," International Conference on Machine Learning and Cybernetics. IEEE, 2008.

[5] D. O. P. Chinnakotla Manoj Kumar, Ranadive Sagar and B. Pushpak, "Hindi to english and marathi to english cross language information retrieval evaluation," in Advances in Multilingual and Multimodal Information Retrieval, pp. 111–118, Springer, 2008.

[6] A. Almeida and P. Bhattacharyya, "Using morphology to improve Marathi monolingual information retrieval," FIRE Working Note, 2008.

[7] S. Subramaniam, A. K. Singh, P. Dasigi, and V. Varma, "Experiments in clir using fuzzy string search based on surface similarity," in Proceeding of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 682–683, ACM, 2009.

[8] N. A. Nasharuddin, "Cross-lingual information retrieval state-of-the-art," electronic Journal of Computer Science and Information Technology (eJCSIT), vol. 2, no. 1, pp. 1–5, 2010.

[9] Nasharuddin, Nurul Amelina, et al, "A review on the cross-lingual information retrieval," Information Retrieval and Knowledge Management, (CAMP), 2010 International Conference. IEEE, 2010.

[10] K. Sourabh, "An extensive literature review on clir and mt activities in india," International

Journal of Scientific and Engineering Research, 2013.
[11] A. Piyush, J. Foster, and G. J. Jones, "Dcu at fire 2013: cross-language !ndian news story search," 2013.
[12] H.B. Patil, A.S. Patil and B.V. Pawar, "Part-of-Speech Tagger for Marathi Language using Limited Training Corpora," International Journal of Computer Applications, 2014
[13] S. Varshney and J. Bajpai, "Improving performance of english-hindi cross language information retrieval using transliteration of query terms," arXiv preprint arXiv:1401.3510, 2014.
[14] Pratibha Bajpai and Dr. Parul Verma, "Cross language information retrieval in indian language perspective," IJRET: International Journal of Research in Engineering and Technology, 2014.
[15] Alessio Bosca, M. Casu, M. Dragoni, and C. D. Francescomarino, "Using Semantic and Domain-Based Information in CLIR Systems," The Semantic Web: Trends and Challenges. Springer International Publishing, 2014.
[16] S. Mayanale and S. Pawar, "Survey of clir and mt systems in marathi language," International Journal of Computational Linguistics and Natural Language Processing, 2015 (in press).
[17] Sandhan [Online]. Available: http://tdil-dc.in/index.php?option=comcontent&view=article&id=66, http://tdil-dc.in/Sandhan/locale.jsp?hi
[18] Fire data collection [Online]. Available: http://www.isical.ac.in/clia/data.html
[19] Cfilt resources [Online]. Available: http://www.cfilt.iitb.ac.in/Downloads.html
[20] Apache lucene core [Online]. Available: http://lucene.apache.org/core/