International Journal of Computer Applications Technology and Research
Volume 4– Issue 7, 579 - 583, 2015, ISSN: 2319–8656

Survey on Indian CLIR and MT systems in Marathi Language

Savita C. Mayanale
Computer Engg Dept., DYPCOE, Akurdi
Savitribai Phule Pune University, Pune,
Maharashtra, India.

S. S. Pawar
Computer Engg. Dept., DYPCOE, Akurdi
Savitribai Phule Pune University, Pune,
Maharashtra, India

Abstract: Cross Language Information Retrieval (CLIR) deals with retrieving relevant information stored in a language different from the language of user's query. This helps users to express the information need in their native languages. Machine translation based (MT-based) approach of CLIR uses existing machine translation techniques to provide automatic translation of queries. This paper covers the research work done in CLIR and MT systems for Marathi language in India.

Keywords: Cross Language Information Retrieval, NLP, Machine Translation, Marathi.

1. INTRODUCTION

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 whereas Cross Language Information Retrieval System (CLIR) retrieves information written in a different language from the query language. 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. As a result CLIR has received more research attention and is increasingly being used to retrieve information on the Internet [1].

Most of the information on the internet is available in English. However, users who don't use English as first language are also significantly high. Non-English users find it difficult in querying information in English. 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. A unique feature of CLIR allows users to query in their native language and provide search result in English. It translates given query into target language then search and provides the most relevant information to the user. This feature of CLIR separates it from any other translating system.

1.1 Translation Types

CLIR system uses two types of translations: Query and Document translation [2]. In Query Translation, the given query will be converted from Native language to Target language and then search operation is performed to get the relevant documents. In Document Translation, all the documents 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 resultant documents in Native language. Among the two, the query translation is easier [3] compared with document translation, because of the size of translation. But, the drawback with query translation is that the given query normally will be short and hence ambiguity problem may arise. As, document Translation is not feasible, most research is based on query translation.

1.2 CLIR Phases

CLIR system involves phases: query pre-processing, translation and disambiguation followed by information retrieval. The pre-processing includes stop words removal and morphological analysis to get the root words. The whole query or some words of the query are transliterated into the target language and sent to the search engine. The process of transliteration refers to expressing a word in one language using the orthography of another language. Document retrieval system involves the use of algorithms for information retrieval and the final stage is to display the results.

2. TERMS IN CLIR SYSTEM

2.1 Machine Translation (MT)

Machine Translation is one of the parts of language processing within Computational Linguistic. Machine Translation (MT) refers to the use of computers to automate some of the tasks or the entire task of translating between human languages. However, the MT system is good tool for CLIR, and actually, if good MT software is available, the CLIR task becomes easier [3]. However, in the case of query translation, the MT approach has not always shown better performance than that of dictionary-based translation. One of the reasons can be short queries are insufficient to provide contextual information for translation.

2.2 Bilingual Dictionary

Bilingual dictionaries are specialized in translating text and words from one language to another. Using a bilingual Machine Readable Dictionary (MRD) is the general approach for CLIR when no commercial MT system with an established reputation is available [3]. In general, most retrieval systems are still based on so called bag-of-words architectures, in which both query statements and document texts are decomposed into a set of words (or phrases) through a process of indexing. Thus translation of a query can be easily done by replacing each query term with its translation equivalents appearing in a bilingual dictionary or a bilingual term list.

2.3 POS Tagger

A Part-Of-Speech Tagger (POS Tagger) reads text and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

Singh J. et al. [4] developed a Marathi POS tagger. The general approach used for development of tagger is statistical method using Unigram, Bigram, Trigram and HMM. It presents a clear idea about all the algorithms with suitable examples. It also introduces a tag set for Marathi language which can be used for tagging Marathi text.

Patil H. B. et al. [5] demonstrated a rule-based POS 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 analysing the linguistic feature of the word, its preceding word, its following word, etc.

2.4 Morphological Analyser

Morphological Analyser is a software component which analyses morphology of given text. It senses or finds out the morphemes of an input word.

A. Muley et al. [6] proposed the morphological analysis for Marathi Language using Ruled Bases Approach. This system has been developed to find a root word of a given word and can be used in Gender Recognition as well.

P. Gawade et al. [7] developed the morphological analyser for Marathi, an inflectional language and also a parsed tree i.e. a grammatical structure. The morphological analyser is combined with statistical POS tagger and Chunker to see its impact on their performance so as to confirm its usability as a foundation for NLP applications.

2.5 Transliteration

It is a process of converting text from one script to another. For example English transliteration of Marathi script "गुजरात" is "Gujarat". There are many standard formats possible for Devanagari-English transliteration viz. ITRANS, IAST, ISO 15919, etc. Transliteration is very useful for converting the named entities (NEs) written in one script to another script in NLP applications like CLIR, Multilingual Voice Chat Applications and Real Time MT.

S. Karimi et al. [8] surveyed key methodologies introduced in the transliteration literature. The approaches are categorized based on the resources and algorithms used and the effectiveness is compared.

P. H. Rathod et al. [9] proposed the named entity transliteration for Hindi to English and Marathi to English language pairs using Support Vector Machine (SVM). The source named entity is segmented into transliteration units and classification of phonetic units is done by using the polynomial kernel function of SVM. The system uses phonetic of the source language and n-gram as two features for transliteration.

2.6 Word sense disambiguation

Word Sense Disambiguation is a process of identifying the most appropriate sense of a word that is used in a given sentence. M. Khapra et al. [10] proposed Domain Specific Iterative Word Sense Disambiguation (WSD) for nouns, adjectives and adverbs in the trilingual setting of English, Hindi and Marathi. The methodology proposed relies on dominant senses of words in specified domains. Starting from monosemous words it iteratively disambiguates bi, tri and polysemous words. Corpus biases for senses are combined with information in Wordnet graph structure to arrive at the sense decisions.

M. Khapra et al. [11] proposed a WSD method that can be applied to a language even when no sense tagged corpora for that language is available. This is achieved by projecting wordnet and corpus parameters from another language to the language in question. The approach is centered on a novel synset based multilingual dictionary and the empirical observation that within a domain the distribution of senses remains more or less invariant across languages.

There are some terms in CLIR such as corpus, stop words, precision, recall, etc. which have same meaning as in traditional information retrieval.

3. LITERATURE SURVEY ON CLIR

3.1 Hindi and Marathi to English Cross Language Information Retrieval

Chinnakotla M. K. et al. [12] proposed a Hindi and Marathi to English CLIR systems as part of CLEF 2007 Ad-Hoc Bilingual task. The system uses query translation approach using bi-lingual dictionaries. First, the input query is pre-processed to identify the root words. Then the words are translated using dictionary. If the words are not found in the dictionary then they are transliterated using a simple rule based approach which uses the corpus to return the 'k' closest English transliteration. Then translated/transliterated words are disambiguated using an iterative page-rank style algorithm. Finally, the disambiguated words are given to the monolingual search engine to get the relevant results. For Hindi, a Mean Average Precision (MAP) achieved is higher than that for Marathi.

3.2 Using Morphology to Improve Marathi Monolingual Information Retrieval

Ashish A. et al. [13] studied the effects of lexical analysis on Marathi monolingual search over the news domain corpus of FIRE-2008. The work also observed the effect of processes such as lemmatization, inclusion of suffixes in indexing and stop words elimination on the retrieval performance. The results show that lemmatization significantly improves the retrieval performance of language like Marathi which is agglutinative in nature. Also, it was observed that indexing of suffix terms, which show spacio-temporal properties, further improved the precision. The effects of elimination of stop words were also observed.

3.3 Sandhan

Sandhan is a project developed under Technology Development for Indian Languages (TDIL) programme [14] with an objective to develop a monolingual search system which will cater tourism domain in five Indian languages viz. Bengali, Marathi, Hindi, Tamil and Telugu. In this project, user has facility to submit query by typing using the INSCRIPT or phonetic layout. On-screen keyboard is also provided to submit query. *Sandhan* has the capability to process the query based on its language and retrieve results from the respective language. Snippets generated for each of the retrieved documents, help the user to understand the context of query terms in that document. 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. An additional UNL based semantic search facility has been provided for Tamil language.

3.4 CLIA

The CLIA (Cross Lingual Information Access) [15] Project is a mission mode project executed by a consortium of academic and research institutions and industry partners. CLIA enables users to enter queries in languages they are fluent in, and uses language translation methods to retrieve documents originally written in other languages. CLIA is an extension of the CLIR paradigm, the objective of which is to introduce additional post retrieval processing to enable users make sense of these retrieved documents. This additional processing takes the form of machine translation of snippets, summarization and subsequent translation of summaries and/or information extraction.

3.5 Marathi-English CLIR

We have presented Marathi-English CLIR in paper [16] 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.

4. MT APPROACHES

MT approaches are classified into three categories: rule-based, knowledge-based and corpus-based.

4.1 Rule-based MT

The Rule Based Machine Translation System takes into account semantic, morphological and syntactic information from a bilingual dictionary and grammar. Based on these rules, it generates the output target language from the input source language by producing an intermediate representation. Rule based system is further classified as Direct MT, Interlingua-based and Transfer –based [17].

In Direct Machine Translation, a direct word by word translation of the input source is carried out with the help of a bilingual dictionary and after which some syntactical rearrangement are made [18].

The Interlingua Approach converts words into an intermediate language IL, which is typically a universal language created for the system to use it as an intermediate for translation into more than one target language [18].

The Transfer based approach uses translation rules to translate the input language to the output language. a dictionary to directly convert source into target whenever a sentence matches one of the transfer rules.

4.2 Knowledge-based MT

It uses the knowledge base that converts the source representation into an appropriate target representation before synthesizing into the target sentence [19]. The basic translation strategy is to extract meaning from the input text in source language, represent this meaning in a language independent semantic representation and then render this meaning in a target language [20].

4.3 Corpus-based MT

In this approach, a bilingual text corpus is trained to get the desired output. The corpus based approach is mainly used in Statistical MT and the Example-Based MT System.

Statistical machine translation is a data-oriented statistical framework for translating text from one natural language to another based on the knowledge and statistical models extracted from bilingual corpora. It requires bilingual or multilingual textual corpora of the source and target language or languages [21].

Example Based Machine Translation System uses previous translation examples to translate from source to target language. EBMT System retrieves examples of existing translation in its example-base and provides the new translation based on that example [17].

5. LITERATURE SURVEY ON MT SYSTEMS FOR MARATHI

This section describes the MT system developments in India for Marathi language.

5.1 Literature Survey of Existing Surveys

The authors in [21] and [22] surveyed various MT systems such as Anglabharti, Anubharti, Shiva and Shakti.

Anglabharti uses pseudo-interlingua approach for translating English to Indian languages. The analysis of English as a source language is done only once and it creates intermediate structure – PLIL (Pseudo Lingua for Indian Languages). The domain of this machine translation system has been public health.

Anglabharti-II uses a generalized example-based (GEB) approach for hybridization with Raw Example-Base (REB). It has provisions for automated pre-editing and paraphrasing, generalized and conditional multi-word expressions as well as recognition of named-entities. The system also contains a ‘failure analysis’ module. The failure analysis module consists of heuristics on speculating the reasons for wrong translation.

Anubharti uses a hybridized example-based machine translation approach. It is a combination of example-based, corpus-based approaches and some elementary grammatical analysis. In Anubharti, the traditional EBMT approach has been modified to reduce the requirement of a large example-base.

Anubharti-II uses Generalized Example-Base (GEB) along with Raw Example-Base (REB) MT approach for hybridization. The combination of example-based approach and traditional rule-based approach is used in this system. The example based approach emulates human-learning process for storing knowledge from past experiences and to be used in future. A shallow chunker is used to fragment the input sentence into small units and then they are matched with a hierarchical example-base.

‘Shiva’ and ‘Shakti’ MT systems are developed jointly by Indian Institute of Science, Bangalore, India, Carnegie Mellon University USA, and International Institute of Information Technology, Hyderabad. Shiva is designed using an Example-based and the system Shakti is designed using combination of rule based and statistical approaches. The rules used for target language generation are mostly linguistic in nature and the statistical approach tries to infer or use linguistic information. Semantic information is also used by some modules in the system. Currently the system is working for three languages (Hindi, Marathi and Telugu).

5.2 Anusaaraka

Anusaaraka project [23] started at IIT Kanpur by Rajeev Sangal is now being continued at IIIT Hyderabad. The purpose of the project was the MT of one Indian language to another Indian language. It is not domain specific but the system has been tested mainly for translating children’s stories. The focus of Anusaaraka was not mainly on MT, but it was on language access between Indian languages. It is currently attempting an English-Hindi machine translation. It uses a Paninian Grammar (PG) and exploits the close similarity of Indian languages.

5.3 UNL Based

Dave S et al. [24] developed a translation system using Universal Networking Language (UNL) as the Interlingua structure. The Universal Networking Language is an international project aimed to create an Interlingua for major human languages. Hindi- UNL, English-Hindi, English-

Marathi, English-Bengali and UNL-Hindi, were also developed using UNL formalism. It is easy to add new language in the system for translation.

5.4 Sampark

A consortium of 11 institutions in India has developed ‘Sampark’ [25], a multipart machine translation system to India Language Machine Translation (ILMT) from Indian Language, funded by TDIL program of Department of Electronics and Information Technology (DeitY), Govt. of India. This program uses Computational Paninian Grammar (CPG) for analyzing language and combines it with machine learning. It is developed using both traditional rules-based and dictionary-based algorithms with statistical machine learning. This consortium has developed language technology for 9 Indian languages resulting in Machine Translation for 18 Indian language pairs.

5.5 Anuvaadaksh

Anuvaadaksh [26] was developed by English to Indian Language MT (EILMT) consortium. Anuvadaksh being a consortium based project has a hybrid approach that is designed to work with platform and technology independent modules. This system has been developed to facilitate the multi-lingual community, initially in the domain-specific expressions of tourism. It integrates four MT Technologies: Tree Adjoining Grammar (TAG) based MT, SMT, Analyze and Generate rules (Anlagen) based MT, Example-based MT (EBMT).

5.6 Google Translator

Google Translate [27] is a free translation service that provides instant translations between 57 different languages. Google Translate generates a translation by looking for patterns in hundreds of millions of documents to help decide on the best translation. By detecting patterns in documents that have already been translated by human translators, Google Translate makes guesses as to what an appropriate translation should be. This process of seeking patterns in large amounts of text is called "SMT".

The TABLE I describe the comparison of different Machine Translation (MT) systems for Marathi language with supported languages, developers of the system and MT approach.

TABLE 1 Comparison of MT Systems for Marathi

MT System	Language	Developer	Approach
Anglabharti (1991)	English to IL	IIT, Kanpur	Interlingua based
Anglabharti II (2004)	English to IL	IIT, Kanpur	Example based
Anubharti (1995) Anubharti II (2004)	Hindi to IL	IIT, Kanpur	Hybrid MT
Anusaaraka (1995)	Punjabi, Bengali, Telugu, Kannada, & Marathi to Hindi.	IIT, Kanpur and University of Hyderabad	Direct MT
UNL based (2003)	Between English, Hindi, and Marathi	IIT, Mumbai	Interlingua based
Shiva and Shakti (2003)	English to {Hindi, Telugu, Marathi}	IISc- Bang, IIIT Hyd, and Carnegie Mellon University	Example based
Sampark (2009)	Among Indian Languages	Consortium of institutions	Rule based + SMT
Anuvadaksh	English to {Hindi, Urdu, Oriya, Bangla, Marathi, Tamil}	EILMT consortium	Hybrid
Google Translate	57 different languages	Google	SMT

6. 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. Machine Translation (MT) is one of the approaches for CLIR system which refers to the use of computers to automate some of the tasks or the entire task of translating between human languages. This paper surveys various developments in CLIR and MT systems, specifically for Marathi language. The work done for both CLIR and MT system for Marathi is in its preliminary stage.

7. ACKNOWLEDGMENT

We would like to thank the publishers, researchers for making their resources available and teachers for their guidance. We would also like to extend a heartfelt gratitude to friends and family members who have inspired, guided and assisted us in performing this work.

8. REFERENCES

[1] Sourabh, Kumar. "An Extensive Literature Review on CLIR and MT activities in India." *International Journal of Scientific & Engineering Research* (2013).

[2] Nagarathinam, A., and S. Saraswathi. "State of Art: Cross Lingual Information Retrieval System for Indian Languages." *International Journal of Computer Applications* 35 (2011).

[3] Nasharuddin, Nurul Amelina. "Cross-lingual Information Retrieval State-of-the-Art." *electronic Journal of Computer Science and Information Technology (eJCSIT)* 2.1 (2010): 1-5.

[4] Singh, Jyoti et al. "Development of Marathi part of speech tagger using statistical approach." *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*. IEEE, 2013.

[5] H. B. Patil, et al., “Part-of-speech tagger for marathi language using limited training corpora," *IJCA Proceedings on National Conference on Recent Advances in Information Technology*, vol. NCRAIT, pp. 33{37, February 2014.

[6] Aditi Muley et al., “Morphological Analysis for a given text In Marathi language”, *International Journal of Computer Science & Communication Network,* Vol 4(1),13-17, 2014.

[7] Gaikwad, Pratiksha Gawade Deepika Madhavi Jayshree, and Sharvari Jadhav Rahul Ambekar. "Morphological Analyzer for Marathi using NLP,"2013.

[8] Karimi, Sarvnaz, Falk Scholer, and Andrew Turpin. "Machine transliteration survey." *ACM Computing Surveys (CSUR)* 43.3 (2011): 17.

[9] Rathod, P. H. et al. "Hindi and Marathi to English machine transliteration using SVM." 2013.

[10] Khapra, Mitesh, et al. "Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting." *Proceedings of International Conference on NLP (ICON 2008), Pune, India*. 2008.

[11] Khapra, Mitesh M., et al. "Projecting parameters for multilingual word sense disambiguation." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009.

[12] Chinnakotla, Manoj Kumar, et al. "Hindi to English and Marathi to English cross language information retrieval evaluation." *Advances in Multilingual and Multimodal Information Retrieval*. Springer Berlin Heidelberg, 2008. 111-118.

[13] Almeida, Ashish, and Pushpak Bhattacharyya. "Using morphology to improve Marathi monolingual information retrieval." *FIRE Working Note* (2008).

[14] *Sandhan*. [Online]. Available: http://tdil-dc.in/index.php?option=com_content&view=article&id=66, http://tdil-dc.in/*Sandhan*/locale.jsp?hi

[15] TDIL Research. [Online]. Available: http://tdil.mit.gov.in/Research_Effort.aspx

[16] Savita C. Mayanale, Ms. S. S. Pawar, "Marathi-English CLIR using detailed user query and unsupervised corpus-based WSD" *Vol. 5 - Issue 6 (June - 2015), International Journal of Engineering Research and Applications (IJERA)* , ISSN: 2248-9622.

[17] Tripathi, Sneha et al. "Approaches to machine translation." *Annals of library and information studies* 57 (2010): 388-393.

[18] Sanyal, Sugata, and Rajdeep Borgohain. "Machine Translation Systems in India." *arXiv preprint arXiv:1304.7728* (2013).

[19] Tomita, Masaru, and Jaime G. Carbonell. "Knowledge-Based Machine Translation, The CMU Approach." (1987).

[20] Bao, Junwei, et al. "Knowledge-based question answering as machine translation." *Cell* 2 (2014): 6.

[21] P. Antony, “Machine translation approaches and survey for indian languages,” *Computational Linguistics and Chinese Language Processing Vol*, vol. 18, pp. 47-78, 2013.

[22] Garje, G. V., and G. K. Kharate. "Survey of Machine Translation Systems in India," *International Journal* (2013).

[23] Bharati, Akshar, et al. "Anusaaraka: overcoming the language barrier in India."*arXiv preprint cs/0308018* (2003).

[24] Dave, Shachi, Jignashu Parikh, and Pushpak Bhattacharyya. "Interlingua-based English–Hindi Machine Translation and Language Divergence." *Machine Translation* 16.4 (2001): 251-304.

[25] Sampark: Machine Translation System among Indian languages (2009) [Online]. Available: http://tdildc.in/index.php?option=com_vertical&parentid=74, http://sampark.iiit.ac.in/

[26] Anuvadaksh. [Online]. Available: http://www.tdil-dc.in/tdildcMain/IPR/Anuvaadaaksh.pdf

[27] Google Translate. [Online]. Available: http://translate.google.co.in/about/intl/en_ALL/