CROSS LANGUAGE INFORMATION RETRIEVAL: IN INDIAN LANGUAGE PERSPECTIVE

Pratibha Bajpai[1], Dr. Parul Verma[2]

[1]*Research Scholar, Department of Information Technology, Amity University, Lucknow*
[2]*Assistant Professor, Department of Computer Science, Amity University, Lucknow*

Abstract

Multilingual information is overflowing on internet these days. This increasing diversity of web pages in almost every popular language in the world should enable the user to access information in any language of his choice. But sometimes it is difficult for a user to write her request in a language which she could easily read and understand. This makes cross-language information retrieval (CLIR) and multilingual information retrieval (MLIR) for Web applications a valuable need of the day. It increases the accessibility of web users to retrieve information in any language while post their queries in their native language. The paper critically analyzes the various researchers work in the area of Indian language CLIR. In this paper we also present our prospective prototype for English to Hindi language CLIR. It will also discuss the issues related to the English to Hindi language translation. We had tested 30 queries manually using suggested prototype and found that the precision level is quite good.

Keywords: *Cross lingual Information Retrieval, Query Translation, Sense Disambiguation, English to Hindi Translation.*

----------***----------

1. INTRODUCTION

A classic IR system accepts the user information need in a form of query and gives back the documents that are relevant to the user need. With the explosion of knowledge on the web, it became necessary to break the language barriers for the monolingual IR systems. This may allow the users of IR systems to give query in one language and retrieve documents in different languages.

IR system, with different source and target language is called CLIR system. Cross-Lingual Information Retrieval (CLIR) translates the user query (given in source language) into the target language, and uses translated query to retrieve the target language documents. The drive for evaluation of monolingual and cross-lingual retrieval systems started with Cross-Language Evaluation Forum (CLEF) in European languages and NTCIR in Chinese-Japanese-Korean languages. It is only in the recent past that the Indian languages have gained importance in evaluation. From 2008, a specific campaign focusing on Indian languages started with the Forum for Information Retrieval Evaluation (FIRE). This resulted in the development of large document collection in some Indian languages like Bangla, Hindi, Marathi and Tamil.

Through our paper we like to provide a brief review of the work done by various researchers in the field of Indian languages for CLIR system. The paper is organized as follows: section 2 illustrates different techniques used for query translation. Comparative analysis of CLIR approaches in Indian languages perspective is discussed in section 3. Section 4 describes our prototype for query translation and sense disambiguation while section 5 draws the conclusion.

2. DIFFERENT TECHNIQUES FOR CLIR

Based on different translation resources, three different techniques have been identified in CLIR: Dictionary based CLIR, Corpora based CLIR and Machine translator based CLIR.

2.1 Machine Translation

Machine translation, in simple terms, is a technique that makes use of software that translates text from one language to another language. But machine translation is not all about substitution of words from one language to another only; rather it also involves finding phrases and its counterparts in target language to produce good quality translations. Machine translation is of three types:

2.1.1 Rule Based Machine Translation

Rule based MT uses linguistic information about source and target language. M. Nimaiti and Y. Izumi (2012) developed Japanese Uighur machine translation system using rule based approach. They propose a word-for-word translation system using subject verb agreement in Uighur. The results aren't positive and there are still some rooms for improvement. In case of Indian languages, R.Rajan et. al.(2009) propose a rule based system for translating English sentences to Malayalam by utilizing dependencies from parser, POS tagger and transfer link rules for reordering and rules for morphology.

2.1.2 Statistical Machine Translation

Statistical machine translation generates translations using statistical methods based on bilingual text corpora. Dan Wu

& Daqing He (2010) conducted a series of CLIR experiments using Google Translate for translating queries. Their results show that with the help of relevance feedback, MT can achieve significant improvement over the monolingual baseline, no matter whether the query length are short or long. Kraaij & Simard(2003) experimentally claim that web can be used for automatic construction of parallel corpus which can then be used to train statistical translation models automatically.

2.1.3 Example Based Machine Translation

Example based MT reads similar examples in the form of source text and its translation from the set of examples, adapting the examples to translate a new input. Sato and Nagao (1990) investigated the problem of example selection by approximate matching of input sentences and example sentences, using a similarity measure based on the syntactic similarity of dependency tree structures of a sentence pair in question and on the word distance of corresponding words, which were predefined in a thesaurus. Sumita *et al.* (1990) looked into example-based translation of Japanese noun phrases of the pattern [N1 *no* N2] into English as [N2 *prep* N1] or [N1 N2], based on a distance measure for the input phrase and example phrase, calculated as a linear weighted sum of the distances of the three sub-parts, each of which is predefined in a thesaurus.

2.2 Dictionary Based CLIR

The most natural approach to cross-lingual IR is to replace each query term with most appropriate translations extracted automatically from Machine Readable Dictionaries (MRD). The translation using bilingual dictionaries is simple but Ballesteros and Croft (1996) and Hull & Grefenstette(1996) claim that it leads to a 40-60% loss in effectiveness as compared to monolingual retrieval. A.Pirkola (2001) asserts that the loss can be due to factors as untranslatable search keys due to limitations in dictionaries, processing of derived or inflected word forms, phrase and compound translation and lexical ambiguity in source and target languages. To handle these problems, researchers have made use of domain specific dictionaries for the dictionary coverage problem(Pirkola, 1998, 1999), Stemming and morphological analysis to handle inflected words(Hull, 1996, Krovetz, 1993; Porter, 1990), POS tagging for phrase translation(Ballesteros & Croft, 1997), corpus based query expansion (Ballesteros & Croft, 1998; Nie et al. , 1999; Sheridan et. al., 1997) and query structuring for the ambiguity problem(Pirkola, 1998, 1999; Sperer & Oard, 2000).

2.3. Corpus Based Cross Lingual Information Retrieval

Corpus based CLIR methods use multilingual terminology derived from parallel or comparable corpora for query translation and expansion. There are two types of corpus:

2.3.1 Parallel Corpus

A parallel corpus is a collection where texts in one language are aligned with their translations in another language. Several systems have been developed to mine large parallel corpora from the web. Wang and Lin (2010) give a method which first identifies a set of seed URLs and crawl candidate bilingual websites. The obtained pages are cleaned and bilingual texts collected to construct comparable corpora. Wang et. al. (2004) exploit the bilingual search result pages obtained from a real search engine as a corpus for automatic translation of unknown query terms not included in the dictionary. They propose a PAT-tree based local maxima method for effective extraction of translation candidates. The approach gives excellent results.

2.3.2 Comparable Corpus

Comparable corpus, on the other hand, consist of texts that are not translations, but share similar topics. They can be, e.g., newspaper collections written in the same time period in different countries. Sadat Fatiha (2011) exploit the idea of using multilingual based encyclopedias such as Wikipedia to extract terms and their translations to construct a bilingual ontology or enhance the coverage of existing ontologies. The method show promising results for any pair of languages. Qian & Meng (2008) expanded Chinese OOV phrase with its partial English translation and submitted to the search engine. The translation of OOV words is mined by preprocessing the snippets obtained to extract the main text from the web page. The strings obtained are sorted by weighted frequency to output the top n translation of OOV phrase. The method proves to obtain the translation with high time efficiency and high precision.

3. COMPARATIVE ANALYSIS OF CLIR APPROACHES FOR INDIAN LANGUAGES

Cross-language retrieval is a budding field in India and the works are still in its primitive state. Table 1 analyzes the performance of various approaches used by the researchers for Indian languages. In many approaches the cross-lingual results are comparable to that of mono-lingual approaches.

Table 1: Critical Analysis of CLIR for Indian Languages

Languages	Translation	Size of test data/ Performance	Specific Features
English to Hindi A.Seetha , S.Das & M. Kumar (2007)	Select first equivalent/ preferred –n/ random nth equivalent/ all equivalents from	6219 hindi document test collection/ performance of strategy 1,2,3,4 are 64.80%, 57.90%, 11.83% and 57.13% of monolingual retrieval	The four strategies are used to test the system performance on the number of equivalents in the query translation by selecting n equivalents from the list of the dictionary.

	Bilingual dictionary		
Tamil to English S. Saraswathi & A. Siddhiqaa (2010)	Machine translation and Ontological tree	200 documents from the domain "festival"/ relevance improves by 40% for English and 60% for Tamil	A generic platform is built for bilingual IR which can be extended to any foreign or Indian language working with the same efficiency.
English to Hindi A. Seetha, S. Das, J. Rana & M. Kumar (2010)	Translation by Shabdanjali dictionary & query expansion by Hindi Wordnet.	Fire 2010 Hindi test collection/ method is not very effective	Query expansion reformulates the initial query by adding some new related words so that query provides a wider coverage than the original query.
English to Malayalam P.L. Nikesh, S.M. Idicula, David Peter (2008)	Bilingual dictionary developed in house.	System proves to be efficient for CLIR	A basic system can be constructed quickly once the linguistic tools become available.
English to Hindi Larkey & Connell (2003)	Probabilistic dictionary derived from parallel corpus	41697 Hindi news articles/ method contributes to effective Hindi retrieval	It combines the ranked lists from the Inquery search and the Language Modeling search to obtain the final ranking of retrieved documents.
English to Hindi & Hindi to English S. Sethuramalingam & V. Varma (2008)	Bilingual Dictionary	English corpus consisted of 125,638 news articles from the Telegraph, Calcutta edition while Hindi corpus consisted of 95215 news articles published in Jagran/ English-Hindi CLIR performance is 58% while Hindi-English CLIR is 25% of the monolingual performance	Disjunctive query formulation using weighted keywords give an overall better performance in both CLIR and Multi Lingual scenario.
Tamil to English	Bilingual dictionary	Web/ The approach used improves the significance of the content retrieved and the overall efficiency of the process	Using summarization techniques and snippet clustering the result closet to user's query is displayed.
Bengali & Hindi to English D. Mandal & P. Banerjee (2007)	Machine Translation using Bilingual dictionary	English news corpus of LA Times 2002 containing 135153 documents/ Map for Bengali-English queries is 7.26 & for Hindi-English queries is 4.77	Queries with named entities provided better results as compared to the queries without named entities implying the importance of a very good bilingual lexicon and transliteration tool in CLIR for Indian languages.
Tamil to English D.Thenmozhi & C. Aravindan (2010)	Machine Translation	Agricultural ontology/ Retrieves pages with MAP of 95%	The system exhibits a dynamic learning approach wherein any new word that is encountered in the translation process could be updated to the bilingual dictionary.
Hindi to English R. Udupa & J. Jagarlamudi (2008)	Probabilistic translation lexicon produced by Statistical Machine Learning	Parallel corpus consisting of 100K sentence pairs from the news domain/ Retrieval performance is about 81% of that of monolingual system	Transliteration mining of OOV words from the document performance whereas date restriction hurts the retrieval performance.
Hindi to telugu to English P.Pingali & V.Verma (2006)	Bilingual Dictionary	English news corpus of LA Times 1995 containing 113005 documents & 56472 documents from Glasgow Herald of 1995/ The system is much robust	Simple techniques such as dictionary lookup with minimal lemmatization such as suffix removal is not sufficient for Indian Languages CLIR.
English to Bangla A.Imam & S.	SMT using parallel corpus	English to Bangla corpus of approximately 29000 sentences/	Improving corpus quality is about 3 times effectual than

Chowdhury (2011)		NIST & BLUE scores (scoring system for evaluating the performance of a Machine Translation System.)are 4.6 and 0.39 which is below the standard	increasing the corpus size for English-Bengali SMT.
Tamil to English Pattabhi R.K Rao and Sobha. L (2010)	Bilingual dictionary and ontology	125638 documents from English news magazine "The Telegraph" / Results are encouraging	The system performs well for queries for which the world knowledge has been imparted.

3.1 Observation

Cross lingual information retrieval for foreign languages like English, French, Chinese etc. has been an appealing area for researchers from long time. But Indian languages have grabbed attention only a decade back. The work done by researchers show mixed results in terms of improvement over monolingual retrieval in Indian language perspective. Anurag Seetha & S. Das (2010) performed translation on Fire 2010 Hindi test collection using Shabdanjali dictionary & query expansion by Hindi Wordnet. The method proved to be ineffective. It is because general dictionaries have low coverage problem. To remove this inefficiency Larkey and Connell (2003) used probabilistic dictionary derived from parallel corpus for English to Hindi translation and achieved effective cross lingual retrieval. Pattabhi R.K Rao and Sobha. L. (2010) found encouraging results by incorporating Bilingual dictionary and ontology.

Other researchers have made use of machine translation for cross lingual retrieval. D.Thenmozhi & C. Aravindan (2010) used MT on agricultural domain and retrieved pages with MAP of 95%. MT systems produce high quality translations only in limited domains and are very expensive too. It involves the cost of creating bilingual dictionary, parallel corpora and the construction and evaluation of MT system. R. Udupa & J. Jagarlamudi (2008) used Probabilistic translation lexicon produced by Statistical Machine Learning while A.Imam & S. Chowdhury (2011) used SMT using parallel corpus for English to Bangla translation.

Parallel or comparable corpora are yet other useful resources for CLIR. Parallel corpora are preferred in CLIR because they provide more accurate translation knowledge but due to their scarcity, comparable corpora are often used in CLIR. The above observation concludes that there is a wide scope of research to improve existing algorithms or developing new one to improve the performance level of CLIR system.

4. PROTOTYPE APPROACH

In this section we propose an approach for cross-lingual information retrieval on the web and briefly discuss the components of the proposed design. The major components of the design are: Preprocessing, Query translation, Word sense disambiguation and Information Retrieval. Before we start discussing the major components of the system, we need to know the grammatical complexities of the two languages.

4.1 Grammatical Complexities of English to Hindi Translation

Hindi and English are morphologically different languages. Translating from poor (e.g. English) to rich (e.g. Hindi) morphology is a tough job and requires deeper linguistic investigation during translation. The major differences are:

(i) The basic word order in Hindi is Subject-Object-Verb (SOV) as against SVO word order in English. But in Hindi, the constituents of a sentence can be freely moved around in the sentence without affecting the core meaning. E.g. the following sentence pair conveys the same meaning with different word order:

राम ने सीता को देखा Ram ne Sita ko dekha

सीता को राम ने देखा Sita ko Ram ne dekhaa

The identity of Ram as the subject and Sita as the object in both sentences comes from the case markers ने (ne – nominative) and को (ko –accusative)

(ii) Unlike English, vowel length and Vowel nasalization are meaningful in Hindi e.g.

(Kam) means 'less' and (Kaam) means 'work'

(Puuch) means 'ask' and (puunch) means 'tail'

(iii) In English, prepositions precede the words to which they relate. In Hindi, such words are called postpositions because they follow the words they govern.

(iv) Hindi is morphologically richer than English. This can be illustrated from following example: The plural-marker in the word "boys" in English is translated as ए (e – plural direct) or ओं (on – plural oblique):

The boys went to school लड़के पाठशाला गये

The boys ate apples. लड़को ने सेब खाये

Future tense in Hindi is marked on the verb. In the following example, "will go" is translated as जायेंगे (jaaenge), with एंगे (enge) as the future tense marker:

The boys will go to school. लड़के पाठशाला जायेंगे

(v) There are no articles in Hindi. Definiteness of a noun is indicated through pronoun, context or word order.

(vi) All nouns in Hindi are either masculine or feminine. This means an arbitrary gender is assigned to the nouns that have a neutral gender in English e.g. 'chair' is a feminine noun and 'door' is a masculine noun in Hindi.

4.2 Preprocessing

The first step in any CLIR system is preprocessing of query terms to speed up the translation process without affecting the retrieval quality. This preprocessing is done using tokenization, stemming and stop word removal.

4.2.1 Tokenization

Tokenization is defined as an attempt to recognize the boundaries between words and isolate those parts of a query which should be translated in the source query.

4.2.2 Stop Word Removal

Stop Words are words which do not contain important significance in Search Queries and hence can be removed from the query to increase search performance. Removing stop words can be done using a list that contains all stop words.

4.2.3 Stemming

It maps all the different inflected forms of a word to the same stem. For languages like English which have weaker inflections, simple stemming algorithms can be used. Such algorithms only remove plural endings. In languages with stronger inflections, suffices are joined to the stem end to end. The advanced stemming algorithm can recognize such multiple endings and remove them in an iterative fashion. Porter stemmer, Snowball stemmer etc. are well known advanced stemming algorithm.

4.3 Query Translation

In Query Translation, the given query is converted from Source language to Target language and the obtained query searches the database to get the documents in Target language. Query Translation often suffers from the problem of translation ambiguity and this problem is amplified due to the limited amount of context in short queries. Query translation can be done using any one technique including machine translation, dictionary based or corpus based method. The techniques have already been discussed in section 2. The query translation is quiet complex while translating English to Hindi query as the two languages are morphologically different from each other. Out of vocabulary (OOV) words are not translated even after morphological analysis. This type of words can be transliterated using the target language alphabet and be added to final queries. Not much work has been done for the translation of these two languages by Indian researchers till date.

4.4 Ambiguity Removal in Translated Query

Ambiguity is a common problem with all natural languages i.e. there exist a large number of words in these languages carrying more than one meaning. For instance, the English noun *plant* can mean *green plant* or *factory* or the word *bank* means *financial institution* or *pool of a river*. The correct sense of an ambiguous word can be selected based on the context where it occurs. This task of automatically assigning the most appropriate meaning to a polysemous word within a given context is called word sense disambiguation. Disambiguation algorithms use a variety of resources and follow different techniques. On the basis of resource utilization and their processing techniques, the disambiguation techniques can be classified as Knowledge Based Methods (resources used are Machine Readable Dictionaries, Thesaurus, Lexicons), Supervised Learning Methods (Naïve Bayesian Classifier, Exemplar Based Classifier, Lazy Boosting Algorithm), Minimally Supervised Methods and Unsupervised Methods.

4.5 Information Retrieval after Query Translation and Ambiguity Removal

The retrieval system presents the user a set of documents that match his query. The retrieval model is of three types: The Boolean, Vector Space and Probabilistic model. In Boolean model, queries are represented as Boolean expressions and only those documents that logically match the query is presented to the user leaving behind those documents that do not match at all. The major drawback with this model is that it only judges documents completely matching or not and does not determines the degree of matching. The other two methods present the ranked list of documents depending on the degree of matching. Vector Space method calculates the degree of matching by calculating the angle between the query vector and each document vector. The Probabilistic model estimates the probability that a document is relevant for the query on the basis of the assumption that the probability depends on the query and the document representation only.

Step by Step Evaluation of CLIR Based on Prototype Approach

The steps of the proposed approach can be explained by considering the following queries:

Query 1: **Hunger Strikes**

Tokenization- Using whitespace between words the tokens obtained from the query are **'Hunger'** and **'Strikes'**

Stop Word Removal- No stop words exist in the above query.

Stemming- Next using Porter stemmer, the inflected tokens are reduced to their base form. After stemming, query becomes **Hunger strike.**

Query translation- The required translation of the query is 'भूख हड़ताल'.

Ambiguity Removal- Since the translated query is unambiguous, so no disambiguation is required.

Precision- The precision of the query is **.83,** where the number of relevant documents is 10 out of top 12 retrieved documents appeared on first page.

Query 2: **Alcohol Consumption in India**

Tokenization- Tokens of the above query are **'Alcohol', 'consumption', 'in' and 'India'.**

Stop word removal- Next stop word '**in'** is removed using stop word list given by MIT. The query now becomes **Alcohol consumption India**

Stemming- Stemming using Porter stemmer returns the query as **Alcohol consumpt India**

Query Translation- Hindi translation of the query is 'भारत में शराब की खपत'.

Ambiguity removal- The Hindi translation **"भारत"** is ambiguous i.e. it has multiple senses. It refers to country India as well as the son of Pandu, a Mahabharat character. The correct sense of a word can be identified based on the context of the query in which it appears using disambiguation algorithm.

Precision- The precision of the query is **1.0,** where the number of relevant documents is 10 out of 10 retrieved documents appeared on first page.

The queries have been preprocessed and translated manually using tools like Potter stemmer, Stop Word list by MIT etc. and received positive results. Based on suggested approach we will formulize an algorithm for English to Hindi language query translation for CLIR.

5. CONCLUSIONS

The respective work with regard to Indian languages has gained impetus in last decade and there is much to be explored in this field. It is quite obvious from the observations that there is still a scope of improvement in the performance level of CLIR. We presume that the proposed prototype system will prove to be competent with other existing systems.

REFERENCES

[1]. Ballesteros, L, and Bruce W Croft, "Phrasal Translation and Query Expansion Techniques for Cross Language Information Retrieval". In: Proceedings of 20th International ACM SIGIR Conference in Research and Development in IR 1997.

[2]. Ballesteros, L., and Croft, W.B. 1998. "Resolving ambiguity for cross-language retrieval." In *Proceedings of SIGIR Conference*, pages 64-71, 1998.

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

[4]. Chinnakotla Kumar Manoj, Ranadive Sagar, Bhattacharyya Pushpak and Damani P. Om "Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007", in the working notes of CLEF 2007.

[5]. David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th International Conference on Research and Development in Information Retrieval, pages 49–57, 1996.

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

[7]. Grefenstette, G. (1998b). The problem of cross-language information retrieval. In Grefenstette (1998a), pages 1-9.

[8]. Hsu Hung Ming, Tsai Feng Ming, and Hsin-Hsi Chen Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison. In : AIRS 2006, LNCS 4182, (2006) 1-13.

[9]. Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. Query Expansion by Mining User Logs IEEE. Transactions on Knowledge and Data Engineering, Vol. 15(4) 2003.

[10]. Hiemstra, D. And De Jong, F. 1999. "Disambiguation strategies for cross-language information retrieval". In Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries. 274-293.

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

[12]. Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. Inf. Process. Manage., 41(3):433-455.

[13]. Nakazawa, S. Ochiai, T. Satoh K., and Okumura A. Cross language Information Retrieval based on Comparable Corpora. In: Proceedings of the first NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition (NTCIR1) 1999.

[14]. Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., and Järvelin, K. (2003). Fuzzy translation of cross-lingual spelling variants. In SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 345-352, New York, NY, USA. ACM.

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

[16]. Prasad Pingali and Vasudeva Varma, "IIIT Hyderabad at CLEF 2007 - Adhoc Indian Language CLIR task". In: CLEF- 2007, Cross Language Evaluation Forum 2007 Workshop at Budapest Hungary.
[17]. Pingali, P., Varma, V., "Hindi and Telugu to English Cross Language Information Retrieval", *Cross Language Extraction Forum(CLEF)*, 2006.
[18]. Sperer, R. and Oard, D. 2000. "Structured query translation for cross-language information retrieval." In Proceedings of the ACM SIGIR Conference. ACM, New York, 2000.
[19]. Seetha Anurag, Das Sujoy, Kumar M., Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method. In: Proceedings of 10th International Conference on Information Technology (ICIT2007). Available at http://doi.ieeecomputersociety.org/10.1109/ICIT.2007.40.
[20]. Thenmozhi, D., and Aravindan, C. Tamil-English Cross Lingual Information Retrieval System for Agriculture Society, International Forum for Information Technology in Tamil Conference, October 2009.