P.Iswarya, Dr.V.Radha / International Journal Of Engineering Research And Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 5, September- October 2012, pp.1036-1043

Cross Language Text Retrieval: A Review

P.ISWARYA*, Dr.V.RADHA**

*(Department of Computer Science, Avinashilingam Institute for Home science and Higher Education for Women, Coimbatore-43)

** (Department of Computer science, Avinashilingam Institute for Home science and Higher Education for Women, Coimbatore-43)

ABSTRACT

The World Wide Web has evolved into a tremendous source of information and it continues to grow at exponent rate. Now a day's web servers are storing different types of contents in different languages and their usage is increasing rapidly. According to Online Computer Library Center, English is still a dominant language in the Web. Only small percentage of population is familiar with English language and they can express their queries in English to access the content in a right way. Due to globalization, content storage and retrieval must be possible in all languages, which is essential for developing nations like India. Diversity of languages is becoming great barrier to understand and enjoy the benefits in digital world. Cross Language Information Retrieval (CLIR) is a subfield of Information retrieval; user can retrieve the objects in a language different from the language of query expressed. These objects may be text documents, passages, audio or video and images. Cross Language Text Retrieval (CLTR) is used to return text documents in a language other than query language. CLTR technique allows crossing the language barrier and accessing the web content in an efficient way. This paper reviews some types of translations carried out in CLTR, ranking methods used in retrieval of documents and some of their related works. It also discusses about various approaches and their evaluation measures used in various applications.

Keywords- Ambiguity, Bilingual, Disambiguation, Monolingual, Multilingual, Transliteration.

1. INTRODUCTION

The number of Web users accessing the Internet become increasing day to day because people can access any kind of required information at anytime. Information Retrieval (IR) mainly refers to a process that the user can find required information or knowledge from corpus including different kinds of information [1]. Information Retrieval is the fact that there is vast amount of garbage that surrounds any useful information; such information should be easily accessible and digestible. With 100 million internet users, India is at 3 rd place globally in usage of internet. Though the network shrank the globe, the language diversification is a great barrier to attain full benefit of the digital life. Hence there is a need to develop a technique like Cross Language Text Retrieval which is used to retrieve text documents in a language other than the user used to specify the query. Therefore Internet is no longer monolingual and non English contents are accessed rapidly. There are three different types of AdHoc CLTR which are as follows:

- Monolingual Information Retrieval System – refers to Information Retrieval System that can find relevant documents in the same language as the query was expressed.
- Bilingual Information Retrieval System that allows you to querying in one language and finding documents in another language.
- Multilingual Information Retrieval System- allows you to make query in one language and able to find documents in multiple languages.

This paper continues to focus on following sections: Section 2 explains about different translation types in CLTR and Section 3 about how these translations can be carried out using different approaches. Section 4 states about various methods used for ranking the results while retrieving the documents. Section 5 deals with previous work in CLTR. Finally conclusion is presented in Section 6.

2. TYPE OF TRANSLATION IN CLTR

The most important issue in Cross Language Text Retrieval is that queries and documents are in different languages. When the user pose a query in one language, either query or document or both translation takes place. These translations are done using free text and controlled vocabulary approaches. The following Fig1 shows the overview of CLTR system.

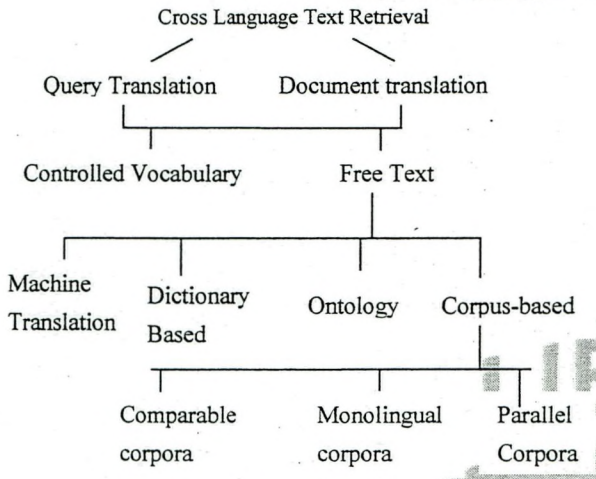

Figure 1: Overview of CLTR system

2.1 Query Translation

Usually the query is translated into the language of target collection of documents. The first step involves parsing the natural language query specified by the user in their native language. The given query sentence is segmented and indexed using Morphological analyzer, Part Of Speech tagging, Stemming and Stop word removal.

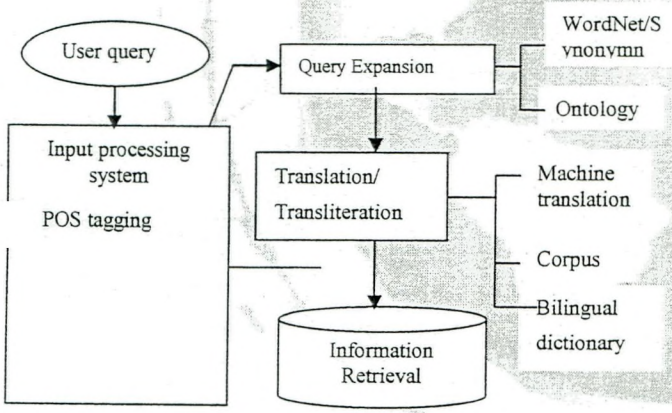

Figure2: Major Steps in Query Translation Approach

Normally user query is short hence ambiguity problem arises; to overcome this drawback the query can be expanded using Word Net/Ontology. Fatiha sadat [2] states that a combination of query expansions before and after query translation will improve the precision of Information Retrieval as well as recall. Translation of query can be done using Machine Translation, Bilingual dictionary and corpus resource. Query translation is simpler than translation of whole documents and it is cost efficient too. The performance of the system heavily depends on how the query is translated effectively. But there are several complexities in achieving good query translation they are translation ambiguity, missing terminology, idioms and compound words and untranslatable terms. The overall process involved in Query translation Process is shown in Fig 2.

2.2 Document Translation

Translation can be done in other way by translating the documents into the language of query and this document translation achieved manually or through various machine translation systems. Translating 400 million non-English web pages of World Wide Web would take 100,000 days (300 years) in one fast PC or 1 month in 3600 PC [3]. However, it is an expensive job to be done once for each query language and most importantly the quality of the translation will be much better because documents provide much more context for a machine translation program to work with. In particular, when it comes to minority languages, the cost becomes almost unbearable.

3. VARIOUS APPROACHES FOR QUERY AND DOCUMENT TRANSLATIONS

As mentioned above the query and document can be translated using the following different techniques such as Controlled vocabulary, free text based or a combination of multiple techniques. The controlled vocabulary is the first and traditional technique widely used in libraries and documentation centres. Documents are indexed manually using fixed terms which are also used for queries. However, this approach remains limited to application whose vocabulary is still manageable. The efficiency and effectiveness degrade radically when size of vocabulary increases.

The alternate approach/way to controlled vocabulary is to use the words which appear in the documents themselves as the vocabulary, such systems are referred as free text retrieval systems.

3.1 Machine Translation approach

Machine Translation (MT) systems that investigates the use of software to translate text or speech from one natural language to another. The main idea of MT system

Table1: Overview of machine Translation system

Systems	Year	Organization/Institute	Domain
Anusaaraka	1995	IIT kanpur	Children stories
Mantra	1999	C-DAC,Bangalore	Rajya sabha Secretariat (official circulars)
Matra	2004	C-DAC,Mumbai	News stories
Angla Bharti	1991	IIT kanpur	Customization

is to carry out a translation without aid of human assistance. Every Machine translation system requires programs for translation, automated dictionaries and grammars to support translation. There are different types of MT exists they are based on direct MT method, Interlingua, transfer method and Empirical machine translation approach. In India machine Translation systems have been developed for translation from English to Indian languages [4] is shown in table 1.

Document translation using MT system is expensive and it is not suitable for large collections and possibly many query languages. Query translation using MT system does not context for accurate translation, it is inadequate for good CLTR.

3.2 Dictionary-based approach

Dictionary approach is used to translate the query, the basic idea in dictionary-based cross language text retrieval is to replace each term in the query with an appropriate term or set of terms in the desired language. CLTR depends on the quality and coverage of dictionary, in manually created bilingual dictionary has good quality but poor coverage. Dictionaries are electronic versions of printed dictionaries and may be general dictionaries or specific domain dictionaries or a combination of both. The major problems of dictionary based approach are translation ambiguity, out-of-vocabulary terms, word inflection and phrase identification. The below Fig 3 shows the overall process in dictionary based translation.

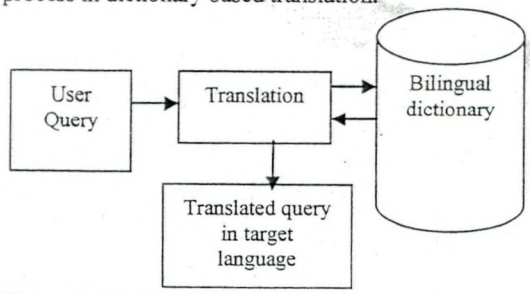

Figure 3: Dictionary based query translation

3.3 Corpus based approach

A Corpus is a repository of a collection of natural language material, it analyzes large collections of existing texts (corpora) and automatically extracts the information needed on which the translation will be based. The aligned corpus contains text samples in one language and their translations into other language are aligned sentence by sentence, word by word, document by document or even character by character. Corpus may contain same language contents from same domain called Monolingual corpora or unaligned corpora.

A parallel corpus is a collection which may contain documents and their translations in more than one language; they are translation-equivalent pairs. Actually a source language query is not translated; it can be matched against the source language component of bilingual parallel corpus. Then target language component aligned to it can be easily retrieved. Parallel corpora can be created using human translation, websites in more than one language or using MT methods. "Spider" systems have been developed to collect documents that have translation equivalents over the internet to produce corpora [5]. The alignment process can be done by comparing documents by the presence of indicators and to construct a parallel corpus is very difficult because they require more formal arrangements. The indicator could be an author name, document date, source, special names in the document, numbers or acronyms, in fact anything which clearly corresponds in both the source and target language texts. Erbug celebi [6] used bilingual parallel corpus consists of 1056 Turkish and 1056 English parallel documents for their experiment sample English to Turkish document shown below table 2.

A comparable corpus is a document collection in which documents are aligned based on the similarity between the topics which they address; they are content-equivalent document pairs. Corpora is hard to maintain, it tend to be domain/application dependent to achieve effective performance.

Table2: Example Document English to Turkish parallel corpus

English Document	Turkish Document
As we are leaving Lefkoşa and heading for Girne, we suddenly decide to climb St. Hilarion castle when our guide cautions us not to leave without seeing it. Leaving behind the area the knights once used for sword practice, we are now in the castle itself, where we have a cold drink and enjoy the view of Girne. As we do so, Mustafa Gürsel, a man with a strong interest in nature and culture in Cyprus, shares with us his knowledge of North Cyprus's endemic plants, the tulip and the orchid.	Lefkoşa'dan ayrılıp Girne'ye doğru yol alırken, rehberimiz "St. Hilarion Kalesi'ni görmeden buradan gitmemelisiniz" deyince hemen kaleye tırmanma kararını veriyoruz. Bir dönem şövalyelerin kılıç alanı olarak kullandığı meydanı geride bırakıp artık kaleye varıyoruz. Kaledeki kafede birşeyler içip aşağıdaki Girne manzarasını izlemeyi ihmal etmiyoruz. Kıbrıs'ın doğası ve kültürüyle yakından ilgili Mustafa (Gürsel) Ağabey de, Kuzey Kıbrıs'ın endemik çiçekleri, lalesi, orkidesi ile ilgili bilgilerini bizimle paylaşıyor bir yandan.

3.4 Ontology-based approach

Ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary. Ontology is an explicit specification of a conceptualization Ontology's can be implemented in translation systems to extract conceptual relations for monolingual and cross language IR.

Sarawathi [7] used Ontological tree for their analysis and keyword retrieval, any number languages can be used without restriction. It requires only single mapping from any language to any other language. It can also used for other purposes such as keywords language identification and sub keyword extraction.

4. RANKING METHODS IN INFORMATION RETRIEVAL

An efficient ranking algorithm is essential in any information retrieval system. The role of ranking algorithms is to select the documents that are most likely be able to satisfy the user needs and bring them in top positions.

Some of the common used ranking methods in cross language information retrieval are discussed below. Michael Sperriosu [8] compared the Okapi BM25 model and Language Modelling (LM) algorithm says that on a very shallow level Okapi outperforms LM. By reviewing simple TFIDF based ranking algorithm may not result in effective CLTR systems for Indian language Queries [9].

4.1 Okapi BM25 model

Consider a user query Q = {q1, q2,....qn} and document D, the BM25 score of the document D is as given in (1) and (2):

$$\text{score}(Q,D) = \sum \text{IDF}(qi).\frac{(f(qi,D).(k1+1))}{f(qi,D)+k1.(1-b+b.(|D|/avgdl))} \quad (1)$$

$$\text{IDF}(qi) = \log .\frac{N - n(qi) + 0.5}{n(qi) + 0.5} \quad (2)$$

where f(qi,D) is the term frequency of qi in D, |D| is length of document D, k1 & b are free parameters to be set, avgdl is the average length of document in corpus, N is the total no. of documents in collection, n(qi) is the number of documents containing qi.

4.2 Language modeling algorithm

The likelihood that a given document d will generate a given query q is used to score candidate documents, and is given in equation (3)

$$p(q|d) = \sum_{w \in q} \omega_w \log \frac{c(w;d) + \mu\delta\, p(w|C)}{\Sigma_w c(w;d) + \mu} \quad (3)$$

where

$w \in q$ is either a term or phrase found in q,

ω_w is the term weight of w given by pseudo relevance feedback (always 1.0 if no pseudo relevant feedback is employed),

$c(w;d)$ is the number of occurrences of w in d,

μ is the Dirichlet prior smoothing parameter,

δ is a parameter used to change the weight of phrases (always 1.0 for non-phrase terms),

$p(w|C)$ is the term count of w in the corpus divided by the corpus size,

and $\Sigma_w c(w;d)$ is the length of document d not including stop-words.

4.3 TFIDF model

The tfidf weight (term frequency–inverse document frequency) is a numerical statistic which reflects how important a word is to a document in a collection or corpus. The term count in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term **t**, within the particular document **d**. Thus we have the term frequency **tf(t,d)**. The inverse document frequency is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient is given by equation (4).

$$\text{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (4)$$

with

|D|: Cardinality of D, or the total number of documents in the corpus

$|\{d \in D : t \in d\}|$: Number of documents where the term t appears (i.e.,tf(t,d)≠0). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the formula to $1 + |\{d \in D : t \in d\}|$

Then the **tf*idf** is calculated using equation (5).

$$tf * idf(t,d,D) = tf(t,d) \times idf(t,D) \quad (5)$$

4.4 Page rank algorithm

Page Rank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. A probability is expressed as a numeric value between 0 and 1. Hence, a PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank. In the general case, the PageRank value for any page 'u' can be expressed in equation (6):

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \quad (6)$$

i.e. the Page Rank value for a page u is dependent on the Page Rank values for each page v contained in the set B_u (the set containing all pages linking to page u), divided by the number $L(v)$ of links from page v.

5. EVALUATION MEASURES

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

5.1 Precision

It is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search is given in equation (7). Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (7)$$

5.2 Recall

It is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents is given in equation (8). Recall in IR is the fraction of the documents that are relevant to the query that are successfully retrieved. It can be looked as the probability that a relevant document is retrieved by the query.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{total relevant documents}\}|} \quad (8)$$

The comparison of various translations and techniques adopted for different languages in different domain are given in table 3.

6. CONCLUSION

In the past five years, research in Cross Lingual Information Access has been vigorously pursued through several international forums, such as, the Cross-Language Evaluation Forum (CLEF), NTCIR Asian Language Retrieval, and Text Retrieval Evaluation Conference (TREC) etc. In this paper, different translations and their various approaches with merits and demerits are stated clearly. The most widely used ranking methods in Information Retrieval are discussed and Cross Language Text retrieval in different languages in different domain is compared. Through investigation of previous research work most of the paper carried out query translation and some researchers have used hybrid approach to achieve query translation and attain acceptable results. In most of research works, document translation is not feasible because of the size of translation. Based on the review, the Okapi BM25 ranking model slightly outperformed LM algorithm and simple TFIDF model is not suitable for Indian language queries. Looking at the statistics of languages on the Internet it seems that there is a huge market for cross-language information retrieval products.

REFERENCES

[1] Tao zhang and Yue-Jie zhang, "Research on Chinese-English CLIR" 2008 *International conference on machine learning and cybernetics, volume-5.*

[2] Fatiha Sadat et al, "Cross language Information Retrieval Via Dictionary-based and Stastistics-based Methods" ,*Conference on Communication, Computer and Signal Processing ,2001.*

[3] Peter schauble, "CLEF 2000 State of Art: Multilingual Information Access", *Eurospider Information Technology AG*

[4] Sanjay Kumar Dwivedi et al," Machine Translation System in Indian Perspectives", *Journal of Computer Science, 2010.*

[5] Mustafa Abusalah.., "Literature Review of Cross language Information Retrieval", *World Academy of science, engineering and technology,2005.*

[6] Erbug Celebi, Baturman Sen, Burak Gunel, "Turkish – English Cross Language Information Retrieval using LSI" *International Symposium on Computer and Information Sciences, 2009.*

[7] Saraswathi et al, "BiLingual Information Retrieval System for English and Tamil", *Journal of Computing, April 2010.*

[8] Michael Speriosu et al, "Comparision of Okapi BM25 ansd Language Modeling Algorithm for NTCIR-6",*Manuscript, Just Systems Corporation, September 2006.*

[9] Sivaji Bandyopadhyay, Tapabrata Mondal, Sudip Kumar Naskar, Asif Ekbal, Rejwanul Haque, Srinivasa Rao

Godavarthy, "Bengali, Hindi and Telugu to English Ad-hoc bilingual task at CLEF 2007 "*Advances in Multilingual and Multimodal Information Retrieval.*

Table 3: Comparison of different translations and their approaches in CLTR.

Authors	Language query	Document language	Domain	Indexing Unit	Translation	Trans-literation	Query expansion	Ranking /Retrieval	Results for CLIR
Tao zhang and Yue-Jie zhang	Chinese and English	English	AP Newswire 88-90	Word based segmentation and n-gram based approach	Bilingual dictionary	No	synonym dictionary	MIT's method, probabilistic method and automatic relevance feedback	MAP E-EIR: 0.3187 C-ECLIR: 0.2833
Anurag Seetha et al	English and hindi	Hindi	Newspapers 2003-2004	No	Shabdanjali bilingual dictionary	No	No	Monolingual retrieval system	Average Precision:Monolingual:0.5318 CLIR:0.3446
Vivek Pemawat et al	English and hindi	English and hindi	Allahabad museum	Stemming and stop word removal	Dictionary database	Hash datastructure mapping	No	Vector based model and Google API(document translation)	Change in the values of precision and recall as number of documents increases.
Sivaji Bandyopadhyay et al	Bengali, Hindi and telugu	English	Los Angeles Times of 2002	Porter Stemmer,n-gram indexing and zonal indexing	Bilingual dictionary	modified joint source-channel model	yes	TFIDF model	The system performs best for the Telugu followed by Hindi and Bengali.
Zhang Xiao-fei et al	English	Chinese	Random chinese web pages	No	Bilingual parallel corpus	No	No	Vector space model	The proposed method outperforms purely dictionary based baseline by 14.8%
Fatiha Sadat et al	French and English	English	TREC volume 1 and 2 collection	Porter stemmer and stop words	Bilingual dictionary, statistic based method and Parallel Corpora.	No	Interactive relevance feedback, a Domain Feedback and similarity thesaurus.	NAMAZU retrieval system	Different combination of query expansion before and after translation with an OR operator shown a best average precision with 99.13% of monolingual performance.
Jagadeesh Jagarlamudi et al	Hindi, tamil,telugu,Bengali and Marathi	English	Los Angeles Times	Stop word removal and porter stemmer	Bilingual statistical dictionary and word by word translation	yes	no	Language Modeling	CLIR performance: 73% of monolingual system
Prasad Pingali et al	Hindi and telugu	English	Los Angeles Times 2002	Lucene Framework	TFIDF algorithm in combination with Bilingual	yes	no	Vector based ranking using a variant of TFIDF	Hybrid Boolean formulation improves ranking of documents

					dictionary			ranking algorithm	
Erbug Celebi et al	Turkish and English	Turkish and English	Skylife Magazine	Porter stemmer and Longest-match stemming algorithm	Parallel corpus	No	No	Latent Semantic indexing	It increases the retrieval performance 3 times when the direct matching is considered.
Mohammed Aljlayl et al	Arabic	English	Standard TREC -7 and 9 collections	Porter stemmer and K-stem algorithm	ALKAFI Machine Translation system	No	No	AIRE search engine	In this method description field is more effective than title and narrative.
Ari Prikola	Finnish and English	English	TREC medicine and health related topics	K-stem and TWOL morphological analyser	Medical dictionary and general dictionary	No	No	INQUERY retrieval system	This method able to achieve performance level of monolingual system, if the queries are structured.
Antony P.J, et al	English	Kannada(target word)	Indian place names	Segmentation,Romanization	Alligned parallel corpus	Sequence labelling method,SVM kernel	no	no	Comparison with Google Indic – Transliteration proposed model gives better results
Pattabhi R.K Rao T and Sobha. L	Tamil and English	English	The telegraph	Tamil Morphological analyzer,Lucene Indexer and porter stemmer	Bilingual dictionary	Statistical method	WordNet and Description Field	Okapi BM25	MAP:0.3980 Recall precision:0.3742
Manoj kumar Chinnakotla et al	Hindi , Marathi and English	English	Los Angles Times 2002	Stemmer and Morphological Analyzer	Bilingual dictionary	Devanagari to English Transliteration	No	Okapi BM25	MAP Monolingual IR:0.4402 Hindi to English:0.2952 & Marathi to English:0.2163
Saravanan et al	Tamil,English, Hindi	English	The telegraph	Porter Stemmer and Alignment Model	Probabilistic lexicon and Parallel Corpora & Bilingual dictionary	Machine Transliteration, Transliteration Similarity Model	No	Language Modeling	MAP Monolingual IR:0.5133 Hindi to Eng:0.4977 & Tamil to Eng:0.4145
B.Manikandan and Dr.R.Shriram	Tamil	English	Random webpages	Stopword removal and Stemmer	Bilingual dictionary	yes	No	Summarization technique	It finds the efficient strategy to implement query translation. In future the algorithm will be tested for more parameters.
R.Shriram and Vijayan Sugumaran	Tamil	English	On sales system	Stop word list and Porter Stemmer	Lexicon and Ontology	no	WordNet	Data mining methods(categorization and aggregation	The proposed approach performs slightly higher than traditional approach.

								)	
S.Thenmozhi and C.Aravindan	Tamil and English	English	Agriculture	Morphological Analyzer, POS tagger	Machine Translation, Bilingual dictionary and Local word reordering	Named Entity Recognizer	WordNet	No	MAP:95% of monolingual system
aswa et al	Tamil and English	Tamil and English	Festival	Tamil morphological analyzer, POS tagger and porter stemmer	Machine Translation	No	Ontology	Page rank	Tamil Increased by 60%. English increased by 40%
Chaware and Srikantha Rao	Hindi, Gujarathi and Marathi	English	Shopping mall	Text to phonetic algorithm	No	Char by char, char to ASCII mapping	No	-	Efficiency depends on minimum number of keys to be mapped.
Nikolaos Ampazis et al	Greek and English	Greek and English	Hellenic history on the Internet website		Parallel corpus	no	no	LSI-SOM	It performs very well on experiments
Michel L.Littman et al	French and English	English and English	Hansard collection 1986-1989(Canadian parliament proceedings)	No	Machine translation-LSI, Crosslanguage-LSI	no	no	LSI	It performs quite well in CL-LSI.