Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base

Rekha Warrier, Sharvari. S. Govilkar
Department of Computer Engineering
PIIT, Navi Mumbai, India

Abstract- Information retrieval is an important activity especially for cross-language environment. When the knowledge is represented by some means/method, it will be easy to retrieve the information. So, to represent knowledge ontology is a rich source, which may give better approach for information retrieval especially for cross language searching. Cross-language information retrieval (CLIR) is a retrieval process in which the user presents queries in one language to retrieve information in another language. CLIR has gained popularity among Information Retrieval (IR) researches in recent years. CLIR is very much needed; especially when the user only knows his/her native language and it may not be possible to process native language all the time. Simple approaches have been developed for CLIR by using multi-lingual dictionary or Word Net. Ontology will be better choice for CLIR, as it covers the entire context and its relationships, which will be helpful for both user and system provider.

Keywords – Cross language information retrieval, Dictionary-based translation,Corpora-based,Machine Translation,Ontology

I. INTRODUCTION

Information retrieval (IR) system aims to retrieve relevant documents to a user query where the query is a set of keywords. CLIR involves the retrieval of documents in a language other than the query language. The growing requirement on the Internet for users to access information expressed in language other than their own has led to Cross Language Information Retrieval (CLIR) becoming established as a major topic in IR.

Cross-language Information Retrieval (CLIR) can be described at an abstract level as the task of retrieving documents across languages. In some sense, the CLIR task represents one extreme case of the so called vocabulary mismatch problem, i.e. the problem that the vocabulary of a user query and the vocabulary of relevant documents can differ substantially. The bag-of-words (BOW) model notoriously suffers from the vocabulary mismatch problem as the different dimensions are inherently orthogonal, thus neglecting relations between different words in the same language as well as across languages. Therefore, the challenging task of retrieving documents to queries in other languages requires models going beyond the traditional bag-of-words model.

The area of Information Access has evolved to perform many sophisticated tasks such as the information retrieval, question answering tasks, summarization, multimedia information retrieval, text mining and clustering and Web information retrieval. Cross-lingual IR has become more important in recent years. The basic idea behind the cross-lingual IR is to retrieve documents in a language (or called as the target language) different from the query language (or the source language) used by the user to develop the query. This may be desirable even when the user is not a speaker of the language used in the retrieved documents.

Now, most people use some type of modern information retrieval system on a daily basis, whether it is Google or some specially created system for libraries. This deals with asking question in one language and retrieving documents in one or more different languages. The variants of the IR are :-

1) **BLIR**(Bi-Lingual Information Retrieval)
2) **CLIR**(Cross-Lingual Information Retrieval) and
3) **MLIR**(Multi-Lingual Information Retrieval).

The ability to search and retrieve information in multiple languages is becoming increasingly important and challenging in today's environment. Consequently, multilingual and cross-lingual (language) information retrieval (MLIR and CLIR) search engines have received more research attention and are increasingly being used to retrieve information on the Internet. CLIR refers to searching, translating and retrieving information in different languages, but mainly between a source language and a target language.

In this report we will concentrate on CLIR. CLIR deals with asking questions in one language and retrieving documents in different language. One approach to CLIR uses different translation approaches to translate queries to documents and indexes in other languages. As queries submitted to search engines suffer lack of context, translation approaches have great problems with resolving query ambiguity. In our approach, we built a multilingual ontology to be used as a translation base for CLIR.

II. CROSS LINGUAL INFORMATION RETRIEVAL

Cross-language information retrieval (CLIR) is a retrieval process in which the user presents queries in one language to retrieve information in another language. CLIR approaches are decomposed into two research fields :- the first is bilingual MRD and machine translation (MT), and the second is concept driven approaches. In dictionary based query translation the query keywords are translated to the target language using Machine Readable Dictionaries (MRD). MRDs are electronic versions of printed dictionaries, and may be general dictionaries or specific domain dictionaries or a combination of both. The major problem in the bilingual dictionary approach is translation ambiguity in addition to problems of word inflection, problems of translating word compounds, phrases, proper names, spelling variants and special terms. MT systems normally attempt to determine the correct word sense for translation by using context analysis. MT is more efficient in document translation as the context is clearer.

Concept driven approaches such as thesauri and multilingual ontologies bridge the gap between the linguistic term and its meaning. A Bilingual Thesaurus groups words with similar meanings in hierarchies (with several levels) of classes and sections and maps them according to their meanings. However, the thesaurus does not include the definition of words. In fact, words in a group are merely related, not synonymous. In our approach we considered developing a bilingual ontology rather than collecting a thesaurus, because we consider ontology as a generalized collection of knowledge that will be used to add a context to search queries by the query expansion, enabling word sense disambiguation.

Knowledge representation plays an important role in almost any domain. It not only gives an exact conclusion but also useful in decision making. There are many ways to represent knowledge, for example, database, taxonomies etc. But the most prominent is ontology. Ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base. According to this definition, the same ontology can be used for building several KBs, which would share the same skeleton. These skeletons can be extended by adding low level sub concepts or high level concepts that cover new areas. Such ontology will give easy and clear understanding of structure of ontology and inference mechanisms will become easier.

III. SYSTEM DESIGN

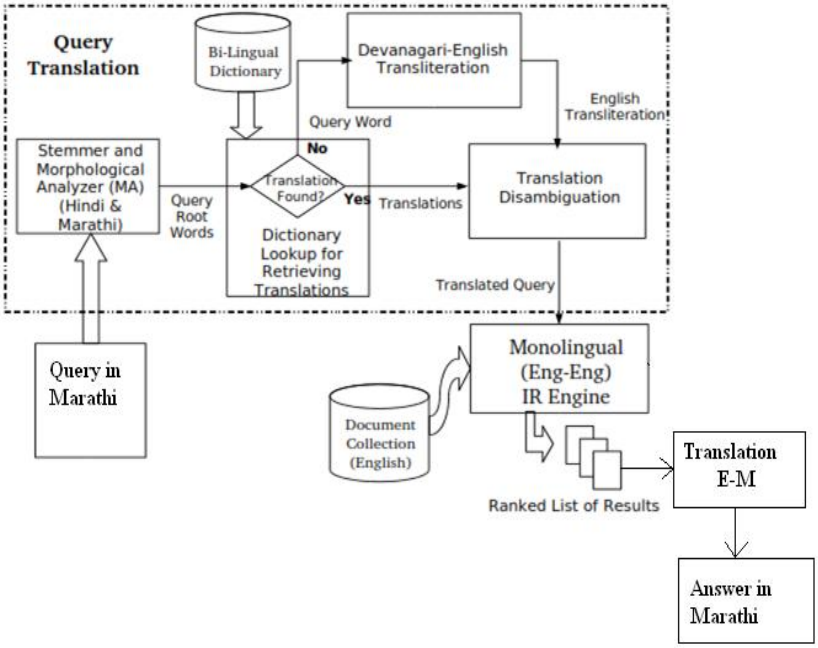

Fig.3: Overall System Architecture

We use a Query Translation based approach in our system since it is efficient to translate the query vis-a-vis documents. It also offers the flexibility of adding cross-lingual capability to an existing monolingual IR engine by just adding the query translation module. We use machine-readable bi-lingual Marathi to English dictionary for query translation. The Marathi to English bi-lingual has less coverage and has around 6110 entries.

Marathi, like other Indian languages, is morphologically rich. Therefore, we stem the query words before looking up their entries in the bilingual dictionary. In case of a match, all possible translations from the dictionary are returned. In case a match is not found, the word is transliterated by the Devanagari to English transliteration module. The above module, based on a simple lookup table and index, returns top three English words from the corpus which are most similar to the source query word. Finally, the translation disambiguation module disambiguates the multiple translations/transliterations returned for the query and returns the most probable English translation of the original query. The translated query is fired against the monolingual IR engine to retrieve the final ranked list of documents as results.
In this architecture, "Devnagari to English Transliteration" and "Translation Disambiguation" are the most important modules other than Stemmer and Morphological Analyzer. This section will give a detailed explanation on the above said modules.

3.1 Devnagari to English Transliteration

Many words of English origin like names of people, places and organizations, are likely to be used as part of the Marathi query. Such words are usually not found in the Marathi to English bi-lingual dictionaries. We use a simple rule based approach which utilizes the corpus to identify the closest possible transliterations for a given Marathi word. We create a lookup table which gives the roman letter transliteration for each Devanagari letter.

Table 3.1 : Lookup table of Roman transliteration

अ a	आ aa/A	इ i	ई I	उ u	ऊ U	ऍ E	ए e	ऐ ai	ऑ O	ओ o	औ au	अं aM	अः a:
क ka	का kaa	कि ki	की kI	कु ku	कू kU	कॅ kE	के ke	कै kai	कॉ kO	को ko	कौ kau	कं kaM	कः ka:
क ka	ख kha	ग ga	घ gha	ङ NGa	च cha	छ Cha	ज ja	झ za	ञ NYa	ट Ta	ठ Tha	ड Da	ढ Dha
ण Na	त ta	थ tha	द da	ध dha	न na	प pa	फ pha/fa	ब ba	भ bha	म ma	य ya	र ra	ल la
व va	श sha	ष Sha	स sa	ह ha	ळ La	क्ष kSha/x	ज्ञ Jha						
क़ Ka	ख़ Kha	ग़ Ga	ज़ Za	ड़ DDa	ढ़ DHa	फ़ Fa	य़ Ya	ऴ LLa	क्र kra	कृ kR	ऋ R		

Since English is not a phonetic language, multiple transliterations are possible for each Devanagari letter. In our current work, we only use a single transliteration for each Devanagari letter. The English transliteration is produced by scanning a Devanagari word from left to right replacing each letter with its corresponding entry from the lookup table. The above approach produces many transliterations which are not valid English words. For example, for the word आस्ट्रेलियाईन (Australian), the transliteration based on the above approach will be astreliyai which is not a valid word in English. Hence, instead of directly using the transliteration output, we compare it with the indexed words in the corpus and choose the 'k' most similar indexed words.

The top 3 closest transliterations for आस्ट्रेलियाईन were australian, australia and estrella. Here we pick the top 3 choices even if our preliminary transliteration is a valid English word. The final choice of transliteration for the source term is made by the translation disambiguation module based on the term-term co-occurrence statistics of the transliteration with translations/transliterations of other query terms.

3.2 Translation Disambiguation

Given the various translation and transliteration choices for the query, the Translation Disambiguation module, out of the various possible combinations, selects the most probable translation of the input query Q. The context within a query, although small, provides important clues for choosing the right translations/transliterations of a given query word. For example, for a query "**नदी जल** " (River Water), the translation for **नदी** is {river } and the translations for **जल** are {water, to burn}. Here, based on the context, we can see that the choice of translation for the second word is water since the combination {river, water}is more likely to co-occur in the corpus than {river, burn}. Consider a query with three words Q = {si,sj,sk}.Let $tr(s_j)= (t_{j1}, t_{j2},\ldots t_{jl})$denote the set of translations and transliteration choices corresponding to a given source word sj where l is the number of translations found in dictionary for sj. The set of possible translations for the entire query Q is T = {tr(si),tr(sj),$tr(s_k)$}. As explained earlier, out of all possible combinations of translations, the most probable translation of query is the combination which has the maximum number of co-occurrences in the corpus.

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

1. User enters the query in Marathi language Q_m.
2. Finds all terms from Q_m and translate those terms into English language using Marathi-English Dictionary and naming them as { t_1 ,t_2,t_3,….t_k}.
3. Finds all terms from Q_m and transliteration those terms into English language using and naming them as {t'_1 ,t'_2 , t'_3 ,…. t_k }
4. Mapping terms { t_1 ,t_2,t_3,….t_k} = {t'_1 ,t'_2 , t'_3 ,…. t_k }
5. Translate Marathi query Q_m into English query Q_E
6. Making all the possible combination of English Query Q_E using{t'_1 ,t'_2 , t'_3 ,…. t_k without replacement of term position up to 2^k times, where k is the number of terms in Q_m
7. Calculate the mean average precision (MAP) of all possible queries that is generated from step 6 and from them choose the best query
8. All the relevant documents generated by the Q_E will be in Q'_E and is again converted into Marathi using bilingual dictionary and above process.
9. Then the answer to the query which was fired will be given back in Marathi to the user.

1.Transliteration:

Algorithm:

i. Once the Marathi query is entered , the tokens/words from the query are separated as below
a. From the start of string, identify all the space positions one by one till the end of string.
b. If space is found then the word before the space / word between two space is identified as a token.
ii. For each token identified in the above step:
a. For each character in the token, its equivalent English character(s) are mapped.
b. We have used below mapping between Marathi alphabets with its equivalent English alphabets.
iii. Once for a complete Marathi query/ string, the tokens are identified and transliterated then the transliterated text will be provided as input to actual translation sub module.

For Example:

The Marathi sentence "शिवाजीचीआईकोणहोती." After stemming and morphological analysis the root words will be शिवाजी , आई , कोण , होत and can be transliterated as "shaivaajai aai kaona haota" in English.

2. Marathi to English Translation module:

Translation involves a change of language altogether. This module takes transliterated output of the transliterated process as a input for Marathi to English translation. Transliterated output will be a Marathi query written in English script. During translation, the transliterate tokens are processed one by one and they are matched for its equivalent English phrase from bilingual dictionary.

Algorithm:

i. Read the transliterated output.
ii. The transliterated tokens are identified from the space positions calculated in the transliterated module.
iii. For each transliterated token, search the bilingual dictionary (M-E) for its equivalent English word match.
iv. Identify query words ('wh' questions) and preorder words (e.g. 'in', 'of' etc.)
v. Organize the English words according to query words and preorder words, this organization need not be grammatically correct, but it will help to generate the node tree.

Example:

Shaivaajai aai kaona haota

The tokenization for the given sentence is:

After tokenization,

A[0]=shaivaajai
A[1]=aai
A[3]=kaona
A[4]=haota

Match each token with bilingual dictionary.

Table 3.2 Tokens

Marathi tokens(after transliteration)	English tokens
shaivaajai	Shivaji
aai, mata, janani	Mother
Kaona	Who
Haota	Was

Output:

who Shivaji mother was (where "who" is query word).

Ontology module provides a knowledge pool in the form of ontology. We are representing ontology in a hierarchical form with sub concepts being related by some relations such as IS-A, Has, of etc. The entire ontology will be traversed to match as concept, sub concept, property or property's value for a query keyword. If found in ontology then extract those ontology terms and make inference from them.

Algorithm: **Retrieve answer from ontology:**

1. Traverse whole ontology
2. First try to match query words in ontology, then match objects and its relation with the given query triples.
3. If found in ontology then extract results from ontology and store as an answer.

Input:

Who(Shivaji,mother)

Output:

Jijabai

3.Translation Module (E-M):

- **Marathi transliteration:**

After the result is fetched from ontology, the Marathi transliteration is done for displaying the answer to the query in a proper sentence. The answer thus we get is a Marathi answer written in the English script.

Algorithm:

1) Maintain the English dictionary for all possible answers to be retrieved from the ontology.
2) Maintain the Marathi dictionary written in the English script where the corresponding Marathi words for the English dictionary words are stored.
3) Take a count of words in the answer retrieved from ontology.
4) For each word search in English dictionary, if found, pick up the corresponding Marathi word from the Marathi dictionary. This way we will get Marathi answer written in English script.
5) We have maintain the query file with the user entered query. Find out the position of "WH" question word from the query and replaced it with the Marathi answer obtained in step (4).
6) In this way we will get Marathi transliterated answer to the query.

For ex:
Input :
jijabai / mother
Output:
Shivaji chi aaijijabaihoti / jijabai Shivaji chi aaihoti

A) Translation (E-M):

Once we get the transliterated Marathi answer, the next and final step is to convert the Marathi transliterated text in Devnagri Marathi script.

Algorithm:

1) We have the Marathi question in Devnagri script stored in query data file
2) Search for the "WH" type of Marathi token from the query data file, say this token as WH- token.
3) Once the WH- token is found the query stored in query data file will be divided into three parts.
i. Pretext before WH-token
ii. WH-token
iii. Post text after WH-token
4) WH-token will be replaced by the answer of query triple (e.g. "Jijabai" in our case).
5) We have answer in English script, which will be converted to Devnagri script as below :
i. Read each English character of the answer and match it with the corresponding Marathi character image.
ii. Whenever vowels are found, the corresponding Marathi strokes will be positioned around the root character.
iii. Repeat step (i) and (ii) till all the English characters of answer are processed.
6) Concatenate the complete answer in Devnagri script as below
i. Pretext before WH-token.
ii. WH-token will be replaced by Devnagri answer obtained in step (5).
iii. Post text after WH-token.

For Ex.
Input:
Shivaji chi aaijijabaihoti
Output:
शिवाजीचीआईजिजाबाईहोती

To understand the above algorithm we shall consider an example.

1) Suppose a user enters a query in Marathi : शिवाजी महाराजांचi जन्म कुठे झाल| ?

So after stemming and morphologically analyzed the query root words will be :

शिवाजी , महाराज , जन्म , कुठे , झाल .

2) These all words will be Q_m . After finding all these terms in Q_m , translate those terms into English language using Marathi-English Dictionary and naming them as

{ t_1 ,t_2,t_3,t_4,t_5}. i.e. {Shivaji , Maharaj , born , where , was }

3) Since some of the translations were not matching, then those terms are put into Devnagari-English Transliteration and name them as { t'_1, t'_2,t'_3,t'_4,t'_5} i.e.

{Shivaji ,Maharaj , janm , kuthe , zaala}

4) Now these two terms are mapped i.e. $t_1=t'_1$;$t_2=t'_2$

{Shivaji,Maharaj,born,where,was}= {Shivaji,Maharaj,janm,kuthe,zaala};

Shivaji = Shivaji; Maharaj=Maharaj; born=janm; where=kuthe; was=zaala

5) Finally translation is stored in Q_E in the same process as explained above

i.e Q_E = {Where was Shivaji Maharaj born }

6) This is returned by Translation Disambiguation and then answer : "Shivaji Maharaj was born at Shivneri" will be the relevant document which will be retrieved from ontology and get stored in Q'_E.

7) This document will be again converted into Marathi using bilingual dictionary and above process and user will get the answer as : शिवाजी महाराजांचi जन्म शिवनेरी येथे झालi

The detailed process of step 6 and 7 are explained above.

IV. CONCLUSION

The Internet has paved opportunities for increasing multi-lingual information exchange and retrieval in future. While this future will be mitigated by the quality of cross lingual search engines to create the connectivity among multi-lingual documents, our study suggest that the current quality of CLIR search engines are poor with scope for much improvement. Creating accurate metadata in different languages in documents or good translation of key information in documents can help improve the quality of the index and retrieval. Cross-lingual IR provides new paradigms in searching documents through myriad varieties of languages across the world and it can be the baseline for searching not only among two languages but also in multiple.

In this experiment, the effectiveness of the ontology based CLIR was better than the Dictionary based one. The benefit of using ontology is not limited to normal word to word translation. These results are especially interesting because they contrast with early monolingual work in which this sort of query expansion degraded rather than improved retrieval effectiveness. It is difficult to determine at this stage whether the improvement is a product of operating in a narrow (and known) domain, the scale and variety of the document collection or some other cause.

After the evaluation of both the pure dictionary and the ontology systems, the ontology based system scored higher in terms of precision. In future development we will enhance and extend the ontology by using annotation tools to align new concepts to the ontology and then test it again with the dictionary system. Other areas for investigation include ease of use, the use of relevance feedback, the effect of more extensive use of concept relations and possibly experiments with larger data sets.

ACKNOWLEDGEMENTS

It is a great pleasure and moment of immense satisfaction for me to express my profound gratitude to my dissertation Project Guide, **Prof. Sharvari Govilkar** whose constant encouragement enabled me to work enthusiastically. Her perpetual motivation, patience and excellent expertise in discussion during progress of the dissertation work have benefited me to an extent, which is beyond expression. I acknowledge all the staff members of the department of Computer and Information Technology of Pillai Institute of Information and Technology for their help and suggestions during various phases of this project work. I would also like to extent my gratitude towards my **family members and friends** who always encouraged me in every deed.

REFERENCES

[1] Mustafa Abusalah,John Tait and Micheal Oakes "*Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base*" September 2009.

[2] F. C. Gey, "*The TEC-2001: Cross Language Information Retrieval Track,"* 2001.

[3] Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin,"*Dictionary-Based CrossLanguage Information Retrieval: Problems, Methods,and Research Finding*" September 2001, Volume 4,pp 209-230.

[4] N.Swapna,Padmaja Rani,Kiran Kumar,"*A Survey on the Cross and MultilingualInformation Retrieval*" National Conference and Research,2012.

[5] Mallamma V Reddy, Dr. M. Hanumanthappa," *Kannada and Telugu Native Languages to English Cross Language Information Retrieval*" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2011, 1876-1880.

[6] Pattabhi R.K Rao and Sobha. L **,**"*Cross Lingual Information Retrieval Track*",AU-KBC Research Centre, MIT Campus, Chennai,2010

[7] Feng YuI, Dequan Zheng and Tiejun Zhao, Sheng Li, Hao Yu, "*Chinese-English Cross-Lingual Information Retrieval based on Domain Ontology Knowledge*",2010

[8] Manoj Kumar Chinnakotla , Sagar Ranadive, Om P. Damani, and Pushpak Bhattacharyya , "Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation ", Department of Computer Science and Engineering, IIT Bombay, India,2008

[9] Sujoy Das1, Anurag Seetha2 , M. Kumar3 and J. L. Rana, "Disambiguation Strategies for English-Hindi Cross Language Information Retrieval System",2009

[10] Dinesh Mavaluru Dr. R. Shriram, " Telugu English Cross Language Information Retrieval: A Case Study ", 2013

[11] Saurabh Varshney1 and Jyoti Bajpai, " Improving performance of English-Hindi cross language information retrieval using transliteration of query terms", 2013

[12] S.M.Chaware , Srikantha Rao, "Ontology Approach for Cross-Language Information Retrieval ",Mumbai,2010

