Tamil-English Cross Lingual Information Translation and Retrieval in Agricultural Domain using VSM

M.SARAVANAN

Assistant Professor, SRM-CARE, SRMIST, Tamilnadu, India

ABSTRACT

Language processing is prompt research area across the country. In that, query translation is one of the major areas of research for the past ten decades. Tamil is morphologically rich and complex language. The suitable morphological processing is very important for Cross Lingual Information Retrieval (CLIR). The contributions towards Tamil to English query translation and transliteration are limited when comparing to other language even though the Tamil language is the official language in south India, Sri Lanka, and Singapore. We propose a CLIR system in agriculture domain for the farmers of Tamil Nadu which helps them to state their query in Tamil and retrieve the documents in English. It uses a bilingual dictionary for translating the user queries and n-gram based approach are used to recover the problem of Word Sense Disambiguation (WSD). The named entities (or) out of vocabulary words are transliterated with the help of statistical system. The retrieved documents are ranked with the help of Vector space model.

Keyword: - *Cross Lingual Information Retrieval, Query transliteration, Query translation, Word sense disambiguation, Vector space model.*

1. INTRODUCTION

Advances in information technology and the wide reach of the internet are radically changing all the activity in our society today. Now a day English is one of World Wide Spoken and Official Language. With the rapid growth distributed system and resources are accessible through the World Wide Web. Cross Lingual Information Retrieval has become complex for the local users to represent, retrieve and understand relevant information from any language. India belongs to several regional spoken languages, due to the lack of communication the people can be misunderstand the theme of document. Tamil peoples have lot of difficulties when they are interacting with the computer system. So the user should aware about the computationally implemented language.

Processing the textual information available in the electronic format and retrieving the information intelligently in responds to users queries have emerged as one of the great challenges in information retrieval. CLIR system is one of the divisions of information retrieval system. A Cross Lingual Information Retrieval process begins when a user enters a query into the system. In CLIR a query does not uniquely identify a single data object in the collection, as an alternative several data objects may match with the query in different degrees of relevancy.

Query translation (QT) is automated translation method which is used to translate a text from one natural language to another using language sources such as bilingual or multilingual dictionaries, WorldNet etc. Normally A word can have 11 morphs (a minimum meaningful unit in a language), A verb can conjugate to 1600 word (a sequence of characters between two successive spaces) forms. The process of identifying the sense of a word in a sentence is called asWSD. WSD is one of the difficult task in translation (both query and document translation). Research has progressed gradually to point out where the WSD method gets sufficient high levels of accuracy.

II. RELATED WORK

Many Research people are working with the CLIR system for different Indian Languages. CLIR system [1] for Indian language (Hindi, Tamil, Telugu, Bengali and Marathi) sub-task of the main Adhoc monolingual and bilingual track. They used Statistical Machine Translation (SMT) with the help of a bilingual dictionary word alignment table.AU-KBC [13] developed a CLIR system for Tamil-English languages by using translation, transliteration and query expansion. Okapi BM25 algorithm is used to provide rank for the retrieved documents.

The morphological analyzer to obtain the core from the query and dynamic learning approach [9] method is used to identify whether the given query word is present in the bilingual dictionary or not. Morphological analyzer[5] is extracting the query words, then it is mapped with the bilingual dictionary to identify whether query word is a root word or not. Machine translation from Tamil to English considered some of the rules sentence reordering. Transliteration [5] is also performed for the named entities present in the query.

Stemmer named as "Maulik" [2] developed for Hindi Language. Stemmer is the process of reducing the inflected words in their base or root form. In Information retrieval techniques stemming act as a significant role to deal with the glossary mismatch problem. The vocabulary words [5] can be translated by using Dictionary based translation method. A single noun [6] can inflected to more than 500 word forms and it includes postpositions. Decision tree classifier [3] is used for the transliteration and modeled as classification that is obtained from Waikato Environment for Knowledge Analysis (WEKA) environment.

Bilingual Translation System [4] for English and Tamil using hybrid approach. Rule Based Machine Translation (RBMT) and Knowledge Based Machine Translation (KBMT) techniques are used for the English to Tamil Translation. If the source query words are present in the bilingual dictionary means then it is allowed to Rule based machine translation and Knowledge based machine Translation. If the query sentence is contains the complex sentence then those complex sentences are split into simple sentences using KBMT and then outcome of KBMT is translated by using RBMT and then processed to get text in target language. If the input sentence is simple then it is directly translated using RBMT.

III. SYSTEM ARCHITECTURE

The CLIR system architecture is illustrated in the Fig.1. It mainly contains the following modules, Text Preprocessing, Verification, Translation, Transliteration, WSD and Retrieval and Ranking.

3.1. Preprocessing module

Text preprocessing module contains the following sequence of process

3.1.1 Tokenization: Tokenization is the process of separating source query words (SQ) into the number of pieces called as tokens. Generally tokens are referred as terms or words, conceivably at the same time throwing away some characters, such as punctuation, question mark, colon, etc. For example,

3.1.2 ***Stop Word Removal****:* In the sentence some words do not help to retrieve the relevant document from the data collection. This kind of words should be removed from the given query is called stop word removal.
Examples of those words such as ஒரு (oru), அந்த (antha), இது (ithu), வரை (varai) etc.Using stop words removal step, these kinds of words are removed from the source query.

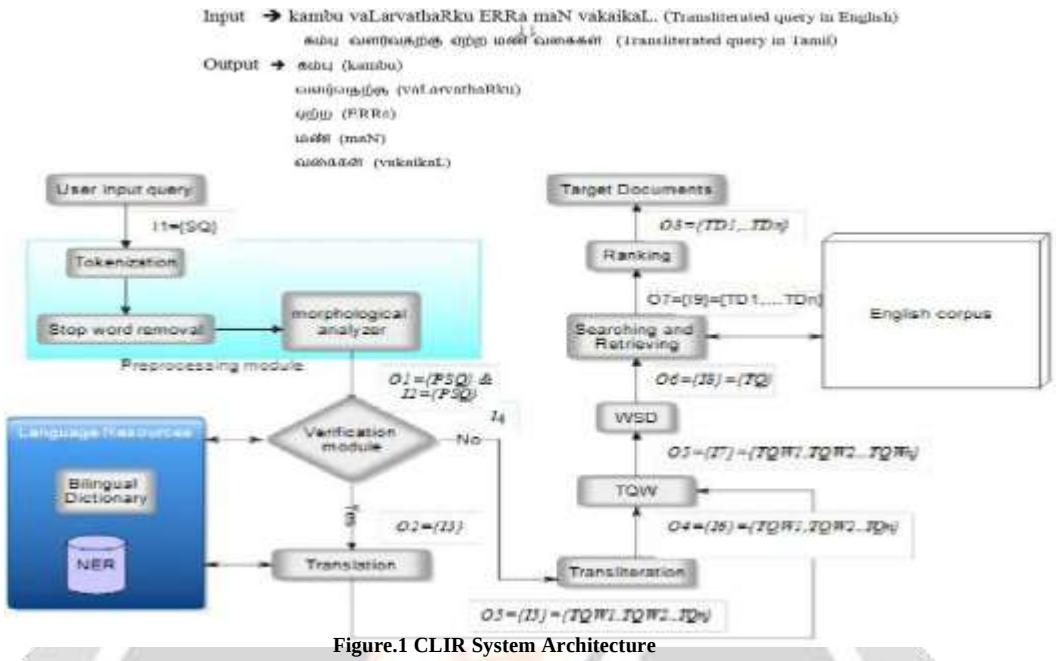

Figure.1 CLIR System Architecture

Input Information Flow

I1={SQ}={Source Language Query}

I2={PSQ}={PSW1,PSW2…PSWn}=
{Preprocessed Source Query Language}

I3={PSW1,PSW2…PSWn}=
{Words found in the bilingual dictionary}

I4={PSW1,PSW2…PSWn}=
{Words not found in the bilingual dictionary}

I5={TQW1,TQW2…TQWn}={Translated Query words}

I6={TQW1,TQW2…TQWn}={Transliterated Query words}

I7={TQW1,TQW2…TQWn}={Target Query words}

I8={TQ}={Target Query}

I9={TD1,TD2,…,TDn}={Target Documents}

Output Information Flow

O1={Preprocessed Source Query}

O2={PSQ}={PSW1,PSW2…PSWn}=
{Preprocessed Source Query Language}

O3={TQW1,TQW2…TQWn}=
{Translated Query words}

O4={TQW1,TQW2…TQWn}=
{Transliterated Query words}

O5={TQW1,TQW2…TQWn}=
{Target Query words}

O6={TQ}={Target Query}

O7={TD1,TD2,…,TDn}={Target Documents}

O8={TD1,TD2,…,TDn}={Target Documents}

3.1.3 ***Morphological Analyzer:*** Morphological analyzer accepts the SQ and it identify the word structure and in the form of morphemes. Morphemes are smallest meaning-bearing units in a language. Morphological analysis primarily deals with the structure of words; morphological analyzer discovers the meaningful constituent morphs in a word. Tamil is one of the agglutinative Dravidian languages, it contains the words by the addition of suffixes to representing various senses or grammatical structure, subsequently the core words or stems. The operation of finite state reinstitution of morphological analyser based on Morphotactics and orthographic rules are described in the Fig.2 & Fig.3.

In this morphological analyzer the noun can fragmented into several circumfix. The possible suffix occurrences in the Tamil noun is given below,

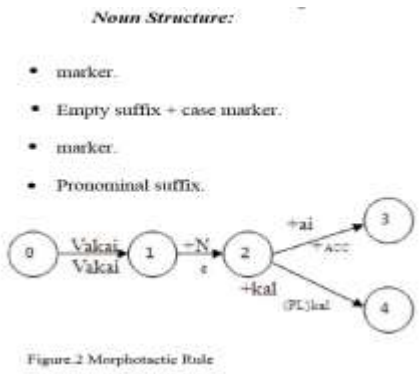

Figure.2 Morphotactic Rule

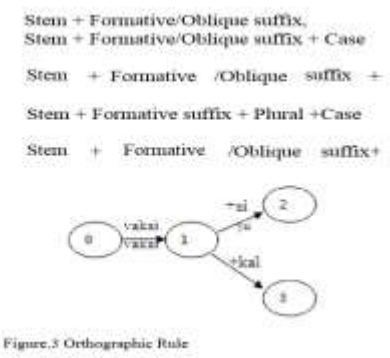

Figure.3 Orthographic Rule

The possible suffix occurrences in the Tamil verbs are given below,

i. ***The finite verb:***

- Root/Stem + Transitive + Causative + Tense / Negative + Empty + PNG Clitics can be added after the Person Number and Gender (PNG) marker.

ii. ***Non -Finite Verb:***

- Root/Stem + Transitive + Causative + Negative + Infinitive / Conditional infinitive suffix.
- Root/Stem + Transitive + Causative + Tense / Negative + Relative Participle / Verbal Participle / Conditional Verbal Participle.
- Root/Stem + Transitive + Causative + Negative + Verbal Noun suffix

The above descriptions mention only the slots on the right side of the root/stem. Exclude this, many non-finite verbs occurs in the left side of the root/stem word and it forms a complex verb structures. Example for complex structure is main + auxiliary verbs. The general structures of the Dravidian verbs are described in the Table.1

<table>
<tr><td rowspan="3">Root/Stem</td><td rowspan="2">Intransitive</td><td>Personal object base</td><td rowspan="2">tense/mode</td><td rowspan="3">Personal endings (person, number and gender)</td></tr>
<tr><td rowspan="2">plural action base</td></tr>
<tr><td>Transitive</td><td>Negative</td></tr>
<tr><td></td><td></td><td>motion base</td><td></td><td></td></tr>
</table>

TABLE.1 THE STRUCTURE OF DRAVIDIAN VERBS

For example, a word ‘கம்பு’ (kambu) &‘மண்’ (maN) consist of single morpheme and ‘வகைகள்’ (vakaikaL) consist of two: the morpheme ‘வகை’vakai and the morpheme‘கள்’ kaL, the morphological analyzer can capable of identify the word ‘vakaikaL’ (வகைகள்)is the plural form of the noun stem ‘vakai’ (வகை).Similarly வளர்வதற்கு (vaLarvathaRku) are separated into two morphemes.i.e. ‘வளர்’ (vaLar) and வ+அதற்கு(v+athaRku), here the (stem word is‘வளர்’(vaLar). The morphological analyzer is not taking care of the naming words. Finally the translated query words are reordered into the SOV (Subject Object Verb).

3.1.4 Verification Module

It is designed for the purpose of verifying the occurrence of PSQ words which are present in Bilingual dictionary or not. After text pre-processing module, the verification module accepts the input query words and performs a dictionary lookup operation to check whether the given query is straightforwardly present in the bilingual dictionary or not. If the words which are found in the dictionary means then it’s given to the Translation Module otherwise it performs the Transliteration operation with the help of Named Entity Resource (NER) if it is predefined named entity.

3.1.5. Dictionary Lookup operation

The Tamil-English bilingual dictionary contains most the words related to agricultural domain. The dictionary had to build as manual and there is no existing resource is available for this domain. After morphological analyzer the stems are given to bilingual dictionary. If it is existing then the meaning of the word is returned. Otherwise the word is then passed on to the subsequent stages in the transliteration module.

3.1.6. Transliteration

Transliteration is defined as 'a transcription from one alphabet to another or to replace the letters or characters of another language with same phonetic sound'. The out of vocabulary words are processed by the Transliteration Module. Transliteration is done using a statistical system based on n-grams approach,N-grams are sequences of characters extracted from a word. A character N-gram is a set of *n* consecutive characters extracted from a word. Typical values for *n* are 2, 3; these correspond to the use of bigrams or trigrams, respectively. Splitting of word is based on the vowels a,e,i,o,u. if the word contains the vowels then it is divided into separate grams. Table.2 illustrates the Transliteration from English to Tamil and Tamil to English. For example, the word 'kambu' results in the generation of the unigrams and bigram, ka+mbu (ka denotes bigrams, mbu denotes trigram).

English to Tamil Transliteration	Tamil to English Transliteration
ka+mbu →க+ம்பு →கம்பு	கம்பு → க+ம்+பு → ka+m+bu → kambu
va+La+r →வ+ள+ர் →வளர்	வளர் →வ+ள+ர் → va+La+r → vaLar
E+RRa → ஏ+ற்ற →ஏற்ற	ஏற்ற →ஏ+ற்+ற→ E+R+Ra→ ERRa
ma+N →ம+ண் →மண்	மண் →ம+ண் → ma+N → maN
va+kai →வ+கை →வகை	வகை→வ+கை → va+kai→ vakai

TABLE II EXAMPLE FOR TRANSLITERATION

3.1.7. Translation

It is depends upon the dictionary based translation method. If the source query words are vocabulary words means system can perform translation process. The source query word is mapped to the bilingual dictionary to check whether it is available or not. If the word is available, then the meaning of the word in target language word (English) is returned. If not, the word is then passed on to the subsequent stages.

3.1.8 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is used to recover the possible senses of a word in the source query. WSD is compared with all possible senses of the nearby words in the given TQW. The count number of words sense is common between calculated and assigned score for the particular sense of the word. The highest score of the sense is declared as the most appropriate one for the target word in the given context. For example, source query "kambu vaLarvathaRku ERRa maN vakaikal" is preprocessed and translated into the target query key words as "peal millet grow soil type" and "stick grow soil type", here "kambu" have ambiguous senses in Tamil. It has two dissimilar senses like "pearl millet" and "stick" get from English Word Net. The first word has highest sense when compared with senses of the other the query words present in the source query. Thus the correct sense of the query word is "pearl millet".

3.1.9. Searching and Retrieval

This module is designed for searching and retrieving relevant target documents for a given TQ. The array-based technique is used for merging translated and transliterated words by each word in the query will be numbered. TQ will retrieve the relevant target documents from the target documents collection. Here, searching process is designed with the help of the Lucene indexing method. Retrieved documents are given to the vector space ranking method to improve the higher relevancy between user queries (Tamil) and retrieved documents (English). Term frequency (*tf*) and inverse document frequency (*idf*) weight is used in the information retrieval and text mining. The weight of a document *i* in term *j* is calculated from the following formula *eqn (1),*

$$w_{i,j} = tf_{i,j} X\ idf_j = tf_{i,j}\ X\ \log\left(\frac{N}{dfj}\right) \quad (1)$$

Where,

N denotes the number of documents in the collection,

idf denotes the inverse document frequency.

If the term weights are calculated then the ranking function is used to measure similarity between the query and document vectors. The similarity between a document Di and a query *Q* is defined in eqn (3.2),

$$sim(Q, D_i) = \frac{\sum_{j=1}^{V} W_{Q,j}\ X\ W_{i,j}}{\sqrt{\sum_{j=1}^{V} w_{Q,j}^2} X \sqrt{\sum_{j=1}^{V} w_{i,j}^2}} \quad (2)$$

Where,

$W_{Q,j}$ is the weight of term j in the query,

$W_{i,j}$ is the weight of j th term in the i th document.

A cosine similarity measure is used to determine the angle between the document and target query words. *J. Result analysis*

The input queries are translated using WSD and retrieved the documents whose precision is better than the precision of the documents retrieved using the Word by Word Translation (WBWT) technique which is showed in the following Table.3.The precision is calculated by using the following formula 3.

$$Precision = \frac{\{Relavant\ document\} \cap \{Retrieved\ documents\}}{\{Retrieved\ documents\}} \quad (3)$$

Transliterated query in English	Transliterated query in Tamil	Query Translation		Precision (%) for retrieved documents	
		Translated keywords using WSD	Translated keywords using WBWT	WSD translation	WBWT translation
kambu vaLarvathaRku ERRa maN vakaikaL	கம்பு வளர்வதற்கு ஏற்ற மண் வகைகள்	Pearl millet grow suitable soil types	Pearl millet stick grow soil types	97	72
ARukaLil uLLa meen vakaikaL	ஆறுகளில் உள்ள மீன் வகைகள்	River fish types	River six fish types	96	62
Udal nalaththiRku ERRa payir vakaikaL	உடல் நலத்திற்கு ஏற்ற பயிர் வகைகள்	Body health suitable crop type	Body health suitable crop type	95	95
Nellayil viLayum payir vakaikaL	நெல்லையில் விளையும் பயிர் வகைகள்	Nellai grow crop type	Paddy grow crop type	92	61

TABLE III Precision of Retrieved Documents

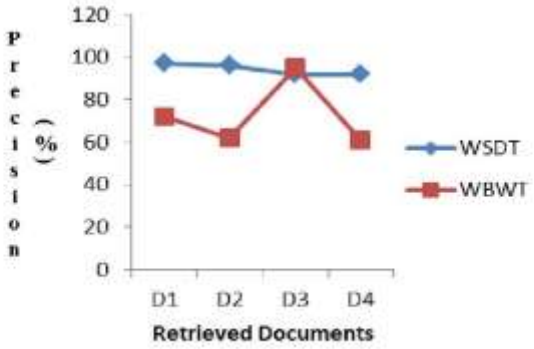

Figure 4 - Comparative study of WSDT & WBWT

V. CONCLUSION

The CLIR System facilitates the Farmers of Tamil Nadu to pose their query in transliterated format to translated target query words. Finally the documents are retrieved from a large corpus in English language. The system focuses

the bilingual dictionary Translation technique to improve WSD rather than the word by word translation. The CLIR systems usually display the search result in English. This system can be additional comprehensive to Rank the pages and provide a summary (in English) of top pages. Document translation is to be considered for final targeted documents in future enhancement.

REFERENCES

[1] PothulaSujatha and P. Dhavachelvan, "A Review on the Cross and Multilingual Information Retrieval", International Journal of Web & Semantic Technology (IJWesT), Vol.2, pp. 115-124, 2011.
[2] Anand Kumar M, DhanalakshmiandV, Soman K.P, "A Sequence Labeling Approach to Morphological Analyzer for Tamil Language", (IJCSE) International Journal on Computer Science and Engineering,Vol. 2, pp. 1944-1951, 2010.
[3] Mallamma V Reddy, Dr. M. Hanumanthappaand Manish Kumar, "Cross Lingual Information Retrieval UsingSearch Engine and Data Mining", ACEEE Int. J. on Information Technology, Vol. 1, pp. 284-293, 2011.
[4] Pattabhi R.K Rao T and Sobha. L, "AU-KBC FIRE 2010 Submission-Cross lingual Information Retrieval Track: Tamil-English", 2010.
[5] C. D Manning, P. Raghavan, and H. Sch¨utze, "An introduction to information retrieval", Cambridge University Press, 2009.
[6] C. J. Van Rijsbergen, "Information retrieval", Butterworths, 1979.
[7] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, "Modern Information Retrieval", ACM Press, 1999.
[8] Cristopher D. Manning. Introduction to information retrieval - cs 276 lecture slides. "In Introduction to Information Retrieval" Stanford University, 2009. http://nlp.stanford.edu/ IR-book/newslides.html.
[9] Pushpak Bhattacharyya, Artificial intelligence - cs 621 lecture slides, IIT Bombay, 2008, http: //www.cse.iitb.ac.in/~pb/cs621.
[10] Kai Gao, Yongcheng Wang, and Zhiqi Wang, "An efficient relevant evaluation model in information retrieval and its application", In International Conference on Computer and Information Technology, volume 0, pages 845–850, Los Alamitos, CA, USA, 2004.
[11] doi:http://doi.ieeecomputersociety.org/10.1109/CIT.2004.1357300.
[12] Vishal Vachhani, Cross-language information access, MTP Stage 1 Report,Indian Institute of Technology,Bombay, 2009.
[13] S.E. Robertson, The Probability Ranking Principle in IR. Journal of Documentation, 33(4):294 – 304, 1977.
[14] K. Sparck Jones, S. Walker, and S. E. Robertson. "A probabilistic model of information retrieval: development and comparative experiments: Part 2. Information Processing & Management", 36(6):809–840, November 2000.
[15] Stephen Robertson (Microsoft Research Cambridge) and Hugo Zaragoza (Yahoo! Research Barcelona), "The probabilistic relevance method: Bm25" and beyond - slides from sigir 2007.
http://www.zaragozas.info/hugo/academic/pdf/tutorial_sigir07_2d.pdf.
[16] A. Singhal," Modern information retrieval: A brief overview", IEEE Data Engineering Bulletin, 24(4):3543, 2001.
[17] S. Brin and L. Page. "The anatomy of a large-scale hyper textual web search engine" Computer networks and ISDN systems, 30(1-7):107–117, 1998.
[18] Ashish Joshi, Kanak Tewari, Ankit Kumar, Bhaskar Pant "A multi-language CLIR with classification by rational agents", International
[19] Conference on Information Systems and Computer Networks, Pages: 94 - 97, DOI: 10.1109/ICISCON.2013.6524181, 2013
[20] V. Klyuev; Y. Haralambous, "Query translation for CLIR: EWC vs. Google Translate", IEEE International Conference on Information
Science and Technology, Pages: 707 - 711, DOI: 10.1109/ICIST.2012.6221738, 2012.

[21] B. Ashwin Kumar, "Profound Survey on Cross Language Information Retrieval Methods (CLIR)", Second International Conference on Advanced Computing & Communication Technologies Pages: 64 -68, DOI: 10.1109/ACCT.2012.91, 2012
[22]. Lan Liu, Yun-Dong Ge, Zhen-Xiang Yan, Jian-Min Yao, "A CLIR-oriented OOV translation mining method from bilingual webpages", International Conference on Machine Learning and CyberneticsVolume: 4, Pages: 1872 - 1877, 2011