Tamil to English Cross Lingual Information Retrieval System for Agricultural Domain Using VSM

M. Saravanan
Department of Computer Science and Engineering, SRM University, Ramapuram Campus, Chennai, Tamilnadu, India.

K. Sathish
Department of Computer Science and Engineering, SRM University, Ramapuram Campus, Chennai, Tamilnadu, India.

M. Prabu
Department of Computer Science and Engineering, SRM University, Ramapuram Campus, Chennai, 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 languages, 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.

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

I. 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 as WSD. 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 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.

A. Preprocessing module

Text preprocessing module contains the following sequence of process

A. 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,
### Input Information Flow

<table>
<thead>
<tr>
<th>I1</th>
<th>Source Language Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>I2</td>
<td>Preprocessed Source Query Language</td>
</tr>
<tr>
<td>I3</td>
<td>Words found in the bilingual dictionary</td>
</tr>
<tr>
<td>I4</td>
<td>Words not found in the bilingual dictionary</td>
</tr>
<tr>
<td>I5</td>
<td>Translated Query words</td>
</tr>
<tr>
<td>I6</td>
<td>Transliterated Query words</td>
</tr>
<tr>
<td>I7</td>
<td>Target Query words</td>
</tr>
<tr>
<td>I8</td>
<td>Target Query</td>
</tr>
<tr>
<td>I9</td>
<td>Target Documents</td>
</tr>
</tbody>
</table>

### Output Information Flow

<table>
<thead>
<tr>
<th>O1</th>
<th>Preprocessed Source Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>O2</td>
<td>Preprocessed Source Query Language</td>
</tr>
<tr>
<td>O3</td>
<td>Translated Query words</td>
</tr>
<tr>
<td>O4</td>
<td>Translated Query words</td>
</tr>
<tr>
<td>O5</td>
<td>Target Query words</td>
</tr>
<tr>
<td>O6</td>
<td>Target Query</td>
</tr>
<tr>
<td>O7</td>
<td>Target Documents</td>
</tr>
<tr>
<td>O8</td>
<td>Target Documents</td>
</tr>
</tbody>
</table>

#### B. 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 `(oru)`, `(antha)`, `(ithu)`, `(varai)`, etc. Using stop words removal step, these kinds of words are removed from the source query.

#### C. 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,

1. **Noun Structure:**
   - Stem + Formative/Oblique suffix,
   - Stem + Formative/Oblique suffix + Case marker.
   - Empty suffix + case marker.
   - Stem + Formative suffix + Plural +Case marker.
   - Pronominal suffix.

![Figure 2: Morphotactic Rule](image)

![Figure 3: Orthographic Rule](image)

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

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

---

**Volume 7 Issue 4 December 2016 283**

**ISSN: 2319 - 1058**
iii. **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 occur 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>
<thead>
<tr>
<th>Root/Stem</th>
<th>Intransitive</th>
<th>Personal object base</th>
<th>tense/mode</th>
<th>Personal endings (person, number and gender)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transitive</td>
<td></td>
<td>plural action base</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>motion base</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For example, a word ‘(kambu)’ (maN) consist of single morpheme and ‘(vakaikaL)’ (vaLarvathaRku) consist of two: the morpheme ‘(vakai)’ (vakai and the morpheme’ kal, the morphological analyzer can capable of identify the word ‘vakaikaL’ (vakaikaL) is the plural form of the noun stem ‘vakai’ (vakai). Similarly (vaLarvathaRku) are separated into two morphemes: i.e. ‘(vaLar)’ (vaLar) and ‘(v+athaRku)’, here the (stem word is ‘(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).

D. **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.

E. **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.

F. **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).

<table>
<thead>
<tr>
<th>TABLE II EXAMPLE FOR TRANSLITERATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root/Stem</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Transitive</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
G. 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.

H. 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 declared as the most appropriate one for the target word in the given context. For example, source query “kambu vaLarvathaRku ERaRa 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”.

I. 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),

\[
W_{ij} = tf_{ij} \times idf_j = \frac{tf_{ij} \times \log \left( \frac{N}{DF_j} \right)}
\]

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 \( D_i \) and a query \( Q \) is defined in eqn (3.2),

\[
similarity(Q, D_j) = \frac{\sum_{j=1}^{m} W_{Q,j} \times W_{D,j}}{\sqrt{\sum_{j=1}^{m} W_{Q,j}^2} \times \sqrt{\sum_{j=1}^{m} W_{D,j}^2}}
\]

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.

\[
\text{Precision} = \frac{\text{Relevant documents}}{\text{Retrieved documents}}
\]
TABLE III Precision of Retrieved Documents

<table>
<thead>
<tr>
<th>Transliterated query in English</th>
<th>Transliterated query in Tamil</th>
<th>Precision (%) for retrieved documents</th>
</tr>
</thead>
</table>
| kamba vara
tahalnka
ERRa maN
vakaikal | கம்பா வாளுதா
ளக்கா
எர்ரா மண்
வாகைகள் | Translated keywords using WSD: Pearl millet grow suitable soil types
Translated keywords using WBWT: Pearl millet stick grow soil types |
| 97 | 72 |
| AKuCaLil uLLa meen vakaikal | ஆகுக்காளில் உல்ல மீன்
வாகைகள் | River fish types 
River six fish types |
| 96 | 62 |
| Udal nalaathi
Kiu
ERRa payir vakaikal | உடல் நல்லாத்தி
கியு
எர்ரா பாயிர்
வாகைகள் | Body health suitable crop type
Body health suitable crop type |
| 95 | 95 |
| Nellaiy vilaiyun
payir vakaikal | எல்லையில்
விலையுன்
பாயிர்
வாகைகள் | Paddy grow crop type
Paddy grow crop type |
| 92 | 61 |

Figure 4 - Comparative study of WSDT & WBWT

IV. 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