Journal of Engineering Science and Technology
Vol. 12, No. 3 (2017) 648 - 666


# ADAPTING HYBRID MACHINE TRANSLATION TECHNIQUES FOR CROSS-LANGUAGE TEXT RETRIEVAL SYSTEM

P. ISWARYA*, V. RADHA

Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women,Coimbatore, India
*Corresponding Author: iswaryacbe333@gmail.com

**Abstract**

This research work aims in developing Tamil to English Cross - language text retrieval system using hybrid machine translation approach. The hybrid machine translation system is a combination of rule based and statistical based approaches. In an existing word by word translation system there are lot of issues and some of them are ambiguity, Out-of-Vocabulary words, word inflections, and improper sentence structure. To handle these issues, proposed architecture is designed in such a way that, it contains Improved Part-of-Speech tagger, machine learning based morphological analyser, collocation based word sense disambiguation procedure, semantic dictionary, and tense markers with gerund ending rules, and two pass transliteration algorithm. From the experimental results it is clear that the proposed Tamil Query based translation system achieves significantly better translation quality over existing system, and reaches 95.88% of monolingual performance.

Keywords: Ambiguity, Hybrid machine translation, Monolingual, Translation.

## 1. Introduction

The Internet is a huge repository of information growing at an enormous rate. India is a multilingual country and it ranks third position globally in the usage of internet. According to Vanopstal et al., around 82% of information provided in WWW is in English and this statistics is increasing in a day-to-day fashion [1]. Similarly, along with this growth, the number of languages used on the web is also diversifying. Non-English speakers are the fastest growing group of new web users and there is a growing interest in non-English sites as the web becomes truly multi-lingual. According to Global Internet Statistics (2004), over 64% of the global web users are non-English speakers.

**Nomenclatures**

| | |
|---|---|
| $BP$ | Brevity Penalty |
| $Q_j$ | Number of relevant documents for Query $j$ |
| $P(doc_i)$ | Precision at $i^{th}$ relevant document |
| $p_n$ | n-gram Precision |
| $x$ | Number of Queries |

**Abbreviations**

| | |
|---|---|
| CLIR | Cross Language Information Retrieval |
| EBMT | Example Based Machine Translation |
| MA | Morphological Analyser |
| MAP | Mean Average Precision |
| MT | Machine Translation |
| NER | Named Entity Recognition |
| OOV | Out of Vocabulary |
| RBMT | Rule Based Machine Translation |
| SMT | Statistical Machine Translation |
| SA | Simulated Annealing |
| AMOSA | Archived Multi Objective Simulated Annealing |
| TDIL | Technology Development for Indian Languages |
| FIRE | Forum of Information Retrieval Evaluation |
| TFIDF | Term Frequency Inverse Document Frequency |
| TQTS | Tamil Query Translation System |
| WSD | Word Sense Disambiguation |

In India, for example, the number of Internet users has crossed the 300 million mark by December 2014 and is expected to reach 500 million users before end of 2016 [2]. Moreover, the Global Reach statistics also shows that nearly 90% of the web users prefer to access the Internet in their native languages [3]. However, most of the users have good reading skills in their native language (large passive vocabulary) but have poor language productive skills (limited vocabulary) in another language (mostly English). Thus they cannot express their information need in non-native language [4, 5] so to solve this issue, the CLIR systems are used.

Cross -Language Information Retrieval (CLIR) system is the solution to cross the language barrier and access the multilingual content in the web. Cross-Language Text Retrieval (CLTR) system is a Sub field of CLIR system, and it allows the user to pose query in one language and retrieve documents in another language. CLTR system involves researchers from the following fields such as Information Retrieval (IR), natural language processing, machine translation and summarization, Speech processing, Document Image Understanding and Human Computer Interaction.

Tamil language is highly agglutinative language and it has been predominantly spoken in south India over 75 million people. This research work develops Tamil to English CLTR system that accepts source query in Tamil language, and retrieves relevant documents in English language. When the user pose a query, either query translation or document translation or both translations should take place [6]. Query translation is simple and cost efficient technique, but its performance heavily depends on how effectively the query is translated. In this research work, enhanced hybrid machine translation technique is implemented in

Tamil to English CLTR system. The Forum for Information Retrieval Evaluation 2011 dataset is used for evaluation purpose.

The paper is organized as follows: Section 2 discusses about translation approaches in CLTR system, and prior works of CLTR are tabulated in Section 3. Section 4 describes about the architecture of the proposed system and its components. In Section 5 experimental results are presented and discussed. Finally conclusion is presented in Section 6.

## 2. Translation Approaches

The CLTR system allows user to supply search queries in the form of text in one's native language, which are then translated and used to retrieve relevant documents in other languages. The translation of queries from one language to another is done using three types of approaches, and they are Knowledge based approach, Corpus based approach and Machine translation approach. The detailed descriptions of these approaches are given below.

### 2.1. Knowledge based approach

Knowledge based approach is divided into three categories such as Thesaurus, Dictionary and Ontology based systems.

#### 2.1.1. Thesaurus based system

The system can be defined as "Controlled Vocabulary" System which represent relationships between terms and concepts that allows user to understand and reformulate better queries. This traditional approach is widely used in commercial and government application centres. Their documents are indexed using fixed terms and that can be used as query terms. It is unsuitable for high volume applications because when the size of indexing vocabulary grows the system becomes unmanageable. Thesaurus based system are costly to build, maintain and mapping between thesauri in different languages is difficult.

#### 2.1.2. Dictionary based system

The dictionary based approach [7, 8] uses a lexical resource to translate words from source language to target document language. The lexical resources are based on knowledge structures, and it is in the form of multi or bi-lingual dictionary. This translation can be done at word level or phrase level. The main assumption in this approach is that user can read and understand documents in target language. In case the user is not conversant with the target language, he/she needs to use some external tools to translate the document in foreign language to their native language. Such tools need not be available for all language pairs. Dictionaries are used to translate each word of the source language query to the desired target language. In the translation process, words can be translated by, not one unique term but a set of terms appearing as equivalent translations in the dictionary. This approach offers a relatively cheap and easily applicable solution

for large-scale document collections. The major problems of dictionary based approach are translation ambiguity, out-of-vocabulary terms, word inflection and phrase identification [6].

### 2.1.3. Ontology based system

Ontology is an explicit specification of conceptualization; it defines the terms and their concepts of the vocabulary in the form of tree. An Ontological tree requires single mapping for translation from one language to any number of languages [9]. It provides better performance than dictionary based system but it is difficult to construct due to its space and time complexity. This system is suitable for specific domain applications and the performance can be improved by automatic construction of ontology from text documents.

## 2.2. Corpus based approach

Corpus is a huge repository collection of textual materials that provides lexical equivalence over languages. The corpus data occurs in two different forms, they are parallel corpus and comparable corpus. Parallel corpora consist of set of documents and their translation equivalents. UN corpus in French, Spanish and English is an example for parallel corpus [10]. Comparable corpora are content equivalent pairs that they have document collections aligned based on their topic, style, and time similarity. An example for comparable corpora is Swiss news agency reports in German, French and Italian [10]. The corpora based system provides high quality results but Indian languages lack for such resources. Also corpora tend to be domain dependent and it has computational complexity.

## 2.3. Machine translation system

Machine Translation (MT) is a task of translating one natural language to another language. But, these systems are able to produce high quality translations only in limited domains [11]. They need information about context and are based on syntactic analysis. Syntactic analysis is not possible for the translation of bag-of-word queries, lacking grammatical structure. However, machine translation has been used as a method in several research reports on cross-language retrieval [12, 13]. The MT system is classified in three paradigms they are rule based, Empirical approach and hybrid based approaches.

### 2.3.1. Rule based MT

Rule Based Machine Translation (RBMT) consists of set of rules created by human experts having linguistic knowledge, aimed at describing the translation process. The traditional MT system can be generalized as Source text analysis, source target transfer and target language generation in conjunction with bi or multilingual dictionaries. A variety of morphological syntactic and semantic information is accumulated and recorded throughout the entire process. RBMT were carried out using three different level approaches and they are Direct, Transfer-based and Interlingua approaches [14].These systems are hard to deal

with ambiguity problem and formulating the rules requires high human involvement. Their rules are universal but they are not domain dependent.

### 2.3.2. Empirical approach

An Empirical approach translates a new incoming text by acquiring knowledge from bilingual parallel corpus. It is also known as Corpus based approach. It is distinguished into two different kinds of approaches such as Example based MT and Statistical MT approach [15].

**Example Based MT (EBMT):** In this approach translation of a new sentence is done by analysing the previously trained example sentences. The EBMT works in four stages namely example acquisition, example base management, and example application and target sentence synthesis. The performance of EBMT depends on the quality of parallel corpus; semantic distance measure and test sentences.

**Statistical MT (SMT):** It is based on statistical models and the model parameters are generated using parallel bilingual corpus. Initially statistical translations models are word based and after significant advancement, phrase based models are introduced. First single word based alignment model is introduced, later it is extended to statistical MT with alignment templates. Various researchers developed efficient search algorithms for an alignment model. SMT systems are learned automatically from the example data and results in faster execution than classical rule based system. But for low resource languages corpus availability is rare, also it is not much suitable for highly different word order languages.

### 2.3.3. Hybrid MT approach

In recent years, hybrid MT approaches has great attention and in some cases translation from source language to target language. They are carried out using rule based approach, followed by statistical approach for adjusting and correcting the output sentences. Pre-processing the input sentences, choosing the best hypothesis and post processing the output data is carried out using rule based and statistical techniques. This technique is better than the previous approaches and has more power, flexibility, and control in translation [14].

## 3. Related works

The related researches that have been done in the field of CLTR system for Indian languages are presented in Table1 as follows:

Through literature review, several CLTR experiments for Indian languages have been carried out using word by word (dictionary) translation approach, but only limited authors have focused on implementing MT techniques in Tamil-English CLTR. An information contained in dictionary and thesaurus based systems are not sufficient and they cannot solve the ambiguity problem which cause significant drop in their performance [30]. Corpus based approach gives good quality translation results, but Indian languages lack in such large corpus that makes the approach unsuitable. There are certain challenges that present in existing word by word translation system [23, 24] such as Out Of Vocabulary

(OOV) terms in bilingual dictionary, ambiguity problem, from short queries gaining of knowledge is limited, named entity handling, and not having proper Part-Of-Speech (POS) tagger in Tamil language. These challenges affect the translation quality and degrade the IR performance. Developing a complete and well specified CLTR models for any language with limited electronic resources is always a challenging and demanding task, where several issues involving translation accuracy and retrieval accuracy are still in the research stage. The performance of the CLTR system depends on the individual performance of each of these steps, and this research work proposes algorithms that aim to improve the working of each of these steps, so as to increase the overall performance of CLTR. . So there is still room for improving the translation and document retrieval process. Machine Translation based systems provides direct resolution of ambiguity in translation by analysing structural and semantic information of source language text. Several researches have attempted to work on hybrid MT [31, 32], by the fact of hybridization techniques that combine the best characteristics of rule and corpus based techniques. Most of the current researches in MT is neither based on purely linguistic knowledge nor on statistics, but includes some degree of hybridization.

**Table 1. Related works of CLTR System for Indian languages.**

| Authors & Year | Query Language | Document language | Domain | Translation | Results for CLIR |
|---|---|---|---|---|---|
| Seetha et al. [16] (2007) | English and Hindi | Hindi | Newspapers 2003-2004 | Shabdanjali bilingual dictionary | Monolingual:0.5318 CLIR:0.3446 |
| Pemawat et al. [17] (2010) | English and Hindi | English and Hindi | Allahabad museum | Dictionary database | Change in the values of precision and recall as number of documents increases. |
| Bandy-opadhyay et al. [18] (2007) | Bengali, Hindi and Telugu | English | Los Angeles Times of 2002 | Bilingual dictionary | The system performs best for the Telugu followed by Hindi and Bengali. |
| Jagarlamudiet al. [19] (2007) | Hindi, Tamil,Telugu,Bengali and Marathi | English | Los Angeles Times | Bilingual statistical dictionary | CLIR performance: 73% of monolingual system |
| Pingali et al. [20] (2007) | Hindi and Telugu | English | Los Angeles Times 2002 | TFIDF algorithm + Bilingual dictionary | Hybrid Boolean formulation improves ranking of documents |

| | | | | | |
|---|---|---|---|---|---|
| Antony et al. [21] (2010) | English | Kannada(target word) | Indian place names | Aligned parallel corpus | Proposed model gives better results than existing. |
| Rao and Devi [ 22] (2010) | Tamil and English | English | The telegraph | Bilingual dictionary | MAP:0.3980 Recall precision:0.3742 |
| Chinnakotlaet al. [23] (2007) | Hindi , Marathi and English | English | Los Angles Times 2002 | Bilingual dictionary | Hindi to English:0.2952 & Marathi to English:0.2163 |
| Saravanan et al. [24] (2013) | Tamil,English, Hindi | English | The telegraph | Bilingual dictionary +Enhanced Transliteration | Hindi to Eng:0.4977 & Tamil to Eng:0.4145 |
| Manikandanand Shriram [25] (2011) | Tamil | English | Random webpages | Bilingual dictionary | It finds the efficient strategy to implement query translation |
| Shriram and Sugumaran [26] (2009) | Tamil | English | On sales system | Lexicon and Ontology | The proposed approach performs better than traditional approach. |
| Thenmozhi and Aravindan [27] (2009) | Tamil and English | English | Agriculture | Statistical Machine Translation | MAP:95% of monolingual system |
| Saraswathi et al. [28] (2010) | Tamil and English | Tamil and English | Festival | Machine Translation, ontological tree | Tamil Increased by 60%. English increased by 40% |
| Chaware and Srikantha [29] (2009) | Hindi, Gujarathi and Marathi | English | Shopping mall | Char by char, char to ASCII mapping | Efficiency depends on minimum number of keys to be mapped. |

The proposed system overcomes the above challenges by insisting improved Tamil POS tagger, Enhancing translation quality using hybrid MT, addition of semantic dictionary with bilingual dictionary, collocation based Word Sense Disambiguation procedure, and use of query expansion technique for short title queries.

## 4. Proposed Methodology

This research work proposes hybrid MT approach, to produce an efficient method that integrates the best features of more than one MT based method, and to compensate for their weakness. The proposed hybrid approach combines rule-based machine translation approach and statistical approach to perform Tamil-English Query Translation. This system is referred to as Tamil Query Translation System (TQTS) in this research, consists of several key tasks that are to be performed in a sequential order to effectively convert the Tamil query to its English equivalent, and retrieving related English documents. These tasks are listed below and the architectural flow is presented in Fig.1.

(i) Tokenization, (ii) Pre-processing, (iii) Translation, (iv) Transliteration and error correction, (v) Query Expansion, (vi) Information Retrieval

### 4.1. Tokenization

The first step of TQTS is tokenization, which is the process of breaking up the query text into units called tokens (words). This process generally use some special symbols like punctuation marks (eg. or -) or spaces as delimiters during word separation. This research work uses blank space as word separator.

### 4.2. Pre-Processing

The pre-processing step of TQTS performs five major tasks, namely, Improved Tamil POS Tagging, Chunking, Named Entity Recognition, Morphological Analysis, and Word Sense Disambiguation. This section presents the proposed algorithmic details of these tasks.

#### 4.2.1. Part-of-Speech tagging

The first step in pre-processing of any language sentence is to retrieve Part Of Speech information that helps in processing many language related activities [33]. POS tagging is defined as a task that reads a set of texts and assigns part of speech label to each of them. As Tamil is highly an inflectional language, for tagging each word, one has to depend on the syntactic function or context to decide upon whether the word is a noun or adjective or adverb or postposition. This leads to a complexity in Tamil POS tagging.

An example of a Tamil sentence along with the POS tagged information is given below.

| | | | | |
|---|---|---|---|---|
| Tamil Sentence: | அவன் | அலுவலகத்தை | நோக்கி | நடந்தான் |
| POS Tagger: | <Proper noun> | <Common noun> | <Interjection> | <Verb Finite> |

Ekbal and Saha [33] proposed a system by extracting 11 features from the annotated corpora with the help of SVM based ensemble method along with an enhanced Simulated Annealing (SA) based Majority Voting Algorithm, Archived Multi Objective Simulated Annealing (AMOSA). It uses an objective function, to increase the accuracies of all the individual POS classes for tagging Bengali and Hindi language words. The accuracy is reduced by more than 12.6% when applied to Tamil language. To increase the accuracy, the present research work introduces a

feature selection algorithm, and enhances the ensemble classifier during POS [34]. An ensemble feature selection combines three algorithms, namely, Split decision tree approach, discriminate function approach and F-score approach, which is initially used to obtain an optimal set of features. The ensemble classifier is improved by hybrid approach using a Wavelet Neural Network (WNN) - Support Vector Machine (SVM). In this SVM-WNN approach, the SVM classifier is used as pre-processor, to reduce the training set to a subset version, by first extracting support vectors, and then WNN is trained using the support vectors.

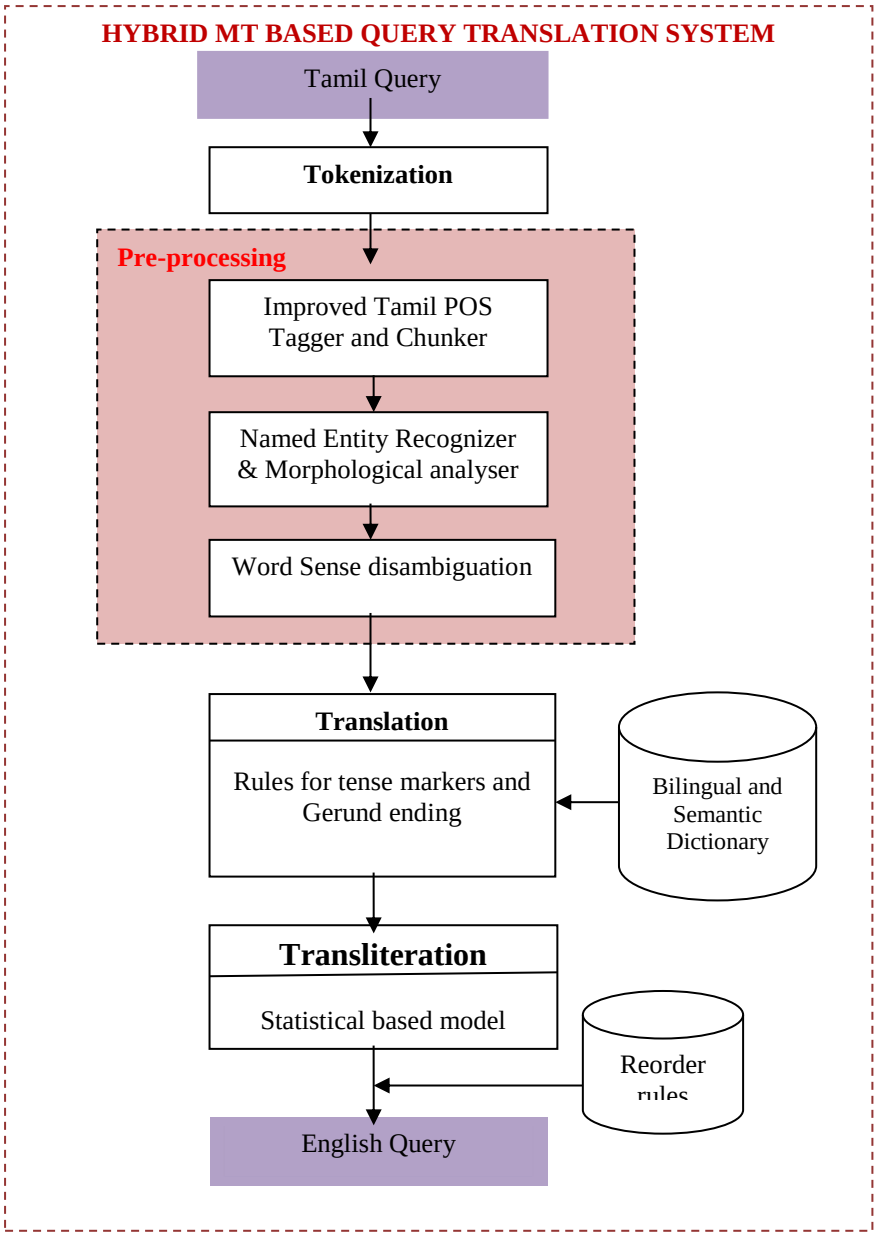


**Fig. 1. Architecture of TQTS.**

### 4.2.2. Chunking

It is a Natural Language Process that separates and segments sentences into their sub constituents such as noun, verb and prepositional phrases. Examples of chunks include noun phrases, prepositional phrases and verb phrases. Chunking works on POS tagged text, so its accuracy depends upon the accuracy of POS tagger. In this research chunking tool is obtained from Technology Development for Indian Languages (TDIL). An example of chunking is shown below.

[அந்த <DET> (B-NP) அழகான <ADJ> (I-NP) பெண் <NN> (I-NP)] NP

### 4.2.3. Named entity recognition

Named entities include the identification of people names, location and companies / organizations, while digits may include time/date stamp and amount. In a Tamil sentence, the NER identifies words that need to be transliterated, and the remaining words are translated using dictionary. In this research work, this is performed using the tool provided by TDIL (http://tdil.mit.gov.in).

### 4.2.4. Morphological analysis

The purpose of an MA is to return root word, and their grammatical information of all the possible word classes for a given word. MA also includes extraction of the grammatical information including number, gender and tense information for all the tokens. As Indian languages have a rich inflectional morphology, MA is an essential tool for such languages. For example consider a word "ஓவியங்கள்", which can be meaningfully divided into ஓவியம் (painting) (noun) + கள் (s) (plural), where the first part represents lexical morpheme and second part is a grammatical morpheme. Machine learning approaches do not require any hand coded morphological rules, and it requires only corpora with linguistical information. These morphological or linguistical rules are automatically extracted from the annotated corpora whereas input is a word and output is root and inflections. In Tamil language, a word may have more than one root word and inflections respectively. In general an input word is denoted as 'W', root and inflections are denoted by 'R' and 'I' respectively ([W] Noun/Verb = [R] Noun/Verb + [I] Noun/Verb). The machine learning classifier used in this research work is SVM-based ensemble classifier [35].

### 4.3.5. Collocation based word sense disambiguation

Word Sense Disambiguation (WSD) algorithm is used to handle collocations. Collocations are defined as nearby words, that strongly suggests the sense of the ambiguous word, in a given occurrence. WSD is an important and challenging task during translation. In general, a WSD algorithm initially uses a manual process to extract collocations, and it identifies sense-collocation words related to the identified collocation using either a dictionary or a thesaurus [36]. This existing process is time consuming, and the manual process may introduce errors. To solve this issue, in this research work, the manual collocation extraction process is replaced using an automatic extraction procedure that uses an enhanced K-Means clustering algorithm.

For example consider a Tamil sentence

Sentence1: **என் கேள்விக்கு விடை** சொல்

Sentence 2: ராமு வீட்டில் இருந்து **விடைப்** பெற்றான்.

In this example, an ambiguous word is "**விடை"** which has at least two possible senses, i.e., answer and relieved. The good quality of translation can only be achieved by choosing a right sense of an ambiguous word, and this process of identifying a correct sense for a word is done using WSD procedure.

The main concern of K-Means algorithm is an optimal selection of 'K' parameter, which is solved using an ensemble approach. An ensemble of clustering algorithms is built with different K values ranging between 2 to 30. This ensemble generates a set of clusters. Majority voting algorithm is then used to find the optimal clustering set from the different partitions created, thus estimating the optimal K value for clustering. The advantage of this approach is that the estimation of this K value is embedded during the process of clustering and requires no extra optimization procedures. The next step uses a sense-collocation dictionary to associate collocations with sense words. The advantage of using automatic extraction step is that it can save search time while considering large number of ambiguous words in a language and reduces manual errors.

## 4.3. Translation

The next step after pre-processing is translation, and it is carried out using knowledge sources and rules set. The outputs of morphological analyser are root word and grammatical morphemes. The root words are directly translated using bi-lingual dictionary. Sometimes several words may not found in the bi-lingual dictionary called OOV words. The problem of OOV is solved in proposed TQTS system with the help of semantic dictionary. If a word is not found in the bi-lingual dictionary, then it is searched in semantic dictionary to obtain an equivalent Tamil word. The semantic equivalent word of the OOV word is translated using root word dictionary. A single word may have several translations, and this ambiguity problem is handled using word sense collocation dictionary. WSD procedure helps in choosing the best hypothesis translation from all possible translations. The remaining part of word belongs to grammatical categories which are translated by applying tense marker and gerund ending rules and some of rules are presented in Fig. 2.

The knowledge based resources such as bilingual dictionary and semantic dictionary are obtained from TDIL (http://tdil.mit.gov.in), and also it collected from various sources of Internet. The transliteration is carried out in next section, and as a result translated English sentence is obtained. Tamil language mostly follows Subject-Object-Verb (SOV) pattern, whereas English language is a Subject-Verb-Object (SVO) pattern. The re-arrangement of Tamil words into the correct structure of English language is done using Rule-based reordering. To rearrange simple sentences from Tamil to English language common reordering rule is applied that is presented below. Finally tagged words are rearranged according to the correct structure of English language.

- If Pronoun (PRP) =" அவன் /அவள்" and VBP="கிறான்/கிறாள்", then add **"is"** after it.
- If Pronoun (PRP) =" அவர்" and VBP="கிறார்கள்", then add **"are"** after it.
- If Pronoun (PRP) =" அவன்" and VBP=" ந்தான்", then add **"was"** after it.
- If Pronoun (PRP) =" அவள்" and VBP=" ந்தாள்", then add **"was"** after it.
- If (PRP) = "அது " and VBD=" கிறது " , then add **"is**" after it.
- If (PRP) = "அது " and VBD=" ந்தது ", then add **"was**" after it.
- If (PRP) = "அவர் " and VBD=" ந்தார்கள் " , then add **"were**" after it.
- If Personal Pronoun (PP) = "நான் " and VBP=" கிறேன்" then add "**am**" after it.
- If (PP) = "நாம் "and VBP= "கிறோம்" then add "**are**" after it.
- If (PP) = "நான் " and VBP= "ந்தேன் " then add **"was**" after it
- If (PP) =" நாம்" and VBP= "ந்தோம்" then add "**were**" after it
- If TO=" க்கு" then add **"to"** before the noun.
- If (PRP) = "அவன்/ அவள் / அவர்/ நான் / நாம்" and MD= "வான்/ வாள்/ வார்கள்/ வேன் / வோம் " then add "will / shall / Could/ Should "respectively.
- If Noun/PRP + ஆல் then add **"with"** before it.
- If Noun/PRP + ஒடு then add **"with"** before it.
- If Noun/PRP + இன் then add **"of"** before it.

**Fig. 2. Tense marker and gerund ending rules.**

**Re-ordering rule from Tamil to English language:** INJ(Interjection) /PP(Personal pronoun) /WP(Wh-pronoun) / WRB(Wh-adverb) / WDT(Wh-determiner) / DT(determiner) / NNP (Proper noun) / PRP (pronoun) / MD(modal) /VBZ (verb present part) /VBP (verb present) /VBN (Verb past part) /VBZ (verb present) /VBD(Verb past ) /VB(verb) /CC (conjunction) /RB (Adverb) /JJ (Adjective) /JJR (Adj-Comparative) /JJS (Adjective-Superlative) /IN (preposition)/TO (to)/NN (Noun)/NNS (Noun plural).

## 4.4. Transliteration with error correction

Transliteration is task of converting one form of script to another form of script. The words that cannot be translated using dictionary are named entities. The NER identify named entities, and give the entities as input to the transliteration engine. Proper nouns and common nouns are often appears in transliterated forms which play an important role in retrieval of documents. Transliteration is first performed to convert named entities and numbers. The first pass retrieval is carried out using a character transformation procedure in which it converts each Tamil character to its English equivalent. For this purpose, a Tamil-English Character Mapping Table (http://www.azhagi.com/az-tamil-modern.html) is used.

During second pass retrieval, statistical transliteration model [24] is implemented that hypothesizes a match between named entity term and a document term in the "comparable" document pair of top 30 retrieval documents (first 30

documents are selected from first pass retrieval). It is an extension of W-HMM word alignment model that makes use of a both the transition and emission models in richer context compared to the classic HMM model.

### 4.5. Query expansion

The Query Expansion is defined as the task of reformulating the translated query by selecting or adding terms to the query, using information obtained from the analysis of the returned documents. The main goal here is to minimize the query-document mismatch and to maximize the retrieval performance. Inclusion of query expansion in CLTR, in general, can improve the retrieval performance by 4-15% [37]. In this research, the method proposed by Lee and Croft [38], is used for query expansion.

### 4.6. Information retrieval

Information retrieval is a process of retrieving relevant documents related to the user query. In this research work Lucene indexer (Lucene is an open source library) is used, which consists of modules for indexing. It is a full-featured text search engine. An Okapi BM25 ranking algorithm *(*en.wikipedia.org/wiki/Okapi_BM25*)* is used to rank the retrieved documents in terms of its relevancy to query words.

## 5. Experimental Results and Discussion

The Forum for Information Retrieval Evaluation (FIRE) 2011 dataset obtained from FIRE organizers to implement adhoc CLTR system which consists of 2, 07,144 documents from September 2005 to December 2010 are used in this experiment. These English news articles are taken from the magazine "telegraph". Articles are from different categories which include sports, business, opinion, stories, front page etc. The FIRE dataset 2011 consists of 50 queries in Tamil, English, Telugu and Hindi languages, and each having a topic, description and narrative, field queries successively which will expand the scope of the query. The sample Tamil queries in FIRE dataset 2011 is presented in Fig. 3.

```
<topics>

<top lang='ta'>
<num>126</num>
<title>பன்றிக் காய்ச்சல் தடுப்பு மருந்து</title>
<desc>இந்தியாவில் தயாரிக்கப்பட்ட கலப்படமில்லாத பன்றிக்காய்ச்சல்
தடுப்பு மருந்து</desc>
<narr>ஆவணங்கள் இந்தியாவில் பன்றி காய்ச்சல் தடுக்க உள்நாட்டு தடுப்பு
மருந்து தயாரிப்பு, மனிதர்கள் மற்றும் விலங்குகள் மீதான அவற்றின்
பயன்பாடு, மருந்து பற்றாக்குறை தடுக்க செய்யப்பட்ட ஏற்பாடுகள், மற்றும்
உயிர் காக்கும் செய்கையில் தடுப்பூசி பங்கு தொடர்பான தகவல்களை
கொண்டிருக்க வேண்டும்.</narr>
</top>
```

**Fig. 3. Sample queries from FIRE dataset 2011.**

Evaluation of the proposed system is done in two stages i) Automatic evaluation of machine translation quality using BLEU score and ii) Tamil to English Cross language text retrieval performance measured using Mean average Precision (MAP) and Precision@10 metrics. BLEU is an automatic evaluation technique which is a geometric mean of n-gram matching. To compute the BLEU score, one has to count the number of n-grams in the test translation that have a match in the corresponding reference translations. IBM's formula for calculating BLEU score [15] is as follows

$$BLEU = BP \times \exp(\sum_{n=1}^{3} \frac{1}{n} \log(p_n)) \quad (1)$$

where brevity penalty is calculated using

$$BP = \min(1, e^{1-\frac{r}{c}}) \quad (2)$$

where c is the length of the corpus of hypothesis translations and r is the effective reference corpus length. The BLEU uses n-gram precision which is termed as $p_n$.

The n-gram precision is calculated as follows

$$p_n = \frac{\sum_{i=1}^{I} \sum ngram \in S_i \quad Count\,(ngram)}{\sum_{i=1}^{I} \sum ngram \in S_i \quad Count_{sys}\,(ngram)} \quad (3)$$

where count (ngram)is the count of n-grams found both in Sentence $s_i$ and reference $r_i$ and count $_{sys}$ (ngram) is the count of n-grams found in $s_i$.

The standard evaluation measures used for this Tamil-English CLTR experiment are Mean Average Precision (MAP) and Precision which are formulated as follows

$$MAP = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(doc_i) \quad (4)$$

MAP is to determine average precision at each query and calculates average for over all queries.

*N* - Number of Queries

$Q_j$ – Number of relevant documents for Query *j*

P ($doc_i$) – Precision at $i^{th}$ relevant document

Precision measure takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n (https://en.wikipedia.org/wiki/Precision_and_recall). Precision is defined as proportion of number of relevant documents retrieved by search to the number of retrieved documents.

$$Precision = \frac{|\{Relevant\ Documents\} \cap \{Retrieved\ documents\}|}{|\{Relevant\ Documents\}|} \quad (5)$$

An automatic Machine Translation based evaluation performance of proposed TQTS system and an existing system is shown in Table 2. Table 3 shows the overall results of Tamil-English CLTR system using FIRE 2011 dataset. The precision and recall curves of monolingual and cross-lingual runs for title, descriptive and narrative queries are presented in Figs. 4(a), (b) and (c) respectively.

**Table 2. MT based Evaluation Statistics of Existing and proposed TQTS system.**

| Type of Queries | BLEU Score for an Existing System | BLEU Score for Proposed TQTS System |
|---|---|---|
| **Title** | 0.7261 | 0.8152 |
| **Descriptive** | 0.6342 | 0.6818 |
| **Narrative** | 0.5014 | 0.5645 |

**Table 3. Comparison between monolingual and cross lingual runs.**

| Queries | Mean Average Precision(MAP) | | Monolingual runs (MAP) | Precision @10 | |
|---|---|---|---|---|---|
| | Existing CLTR System | Proposed TQTS CLTR System | | Existing CLTR System | Proposed TQTS CLTR System |
| **Title** | 0.4962 | 0.5684 | 0.5845 | 0.3 | 0.8 |
| **Descriptive** | 0.5521 | 0.5832 | 0.6176 | 0.5 | 0.7 |
| **Narrative** | 0.5622 | 0.6062 | 0.6312 | 0.5 | 0.8 |

An existing system uses word by word translation approach [22, 24] whereas proposed system implements hybrid MT based approach for Tamil-English CLTR. Table 2 shows that the proposed TQTS system gives higher BLEU score in all three types of queries. Based on the BLEU score value the proposed system provides efficiency gain of 10.92%, 6.98%, 11.17% improvement over word by word approach for title, description and narrative queries respectively. In turn higher translation quality gives more relevant document retrieval against the query.

From Table 3, the cross-lingual performance of proposed and existing system over monolingual run is 84.8%, 97.24% for title queries, 89.39%, 94.24% for descriptive queries and 89.06%, 96% for narrative queries respectively. Query expansion technique is implemented only for short title queries, and it gives greater MAP score value compared to monolingual runs for some queries. But use of Query expansion technique in descriptive and narrative queries, may result in irrelevant terms and irrelevant documents, that causes decline in performance of MAP. Thus the proposed hybrid Machine translation system achieves better quality in translating Tamil to English queries when compared to existing system, and also provides comparable monolingual performance.

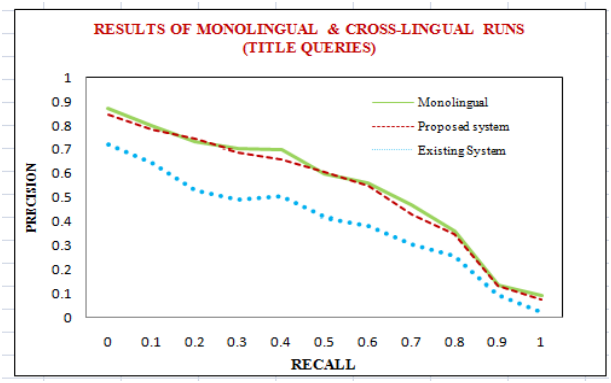


**Fig. 4(a). Precision and Recall curve for Title queries.**

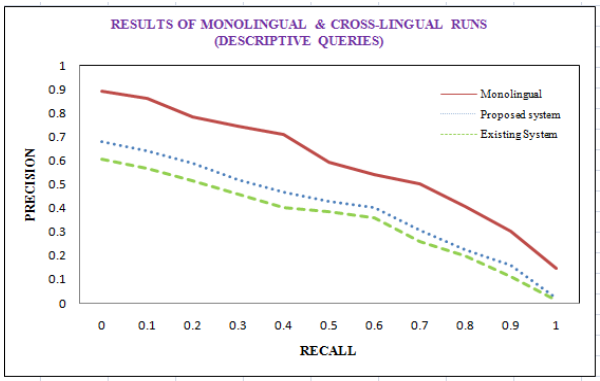


**Fig. 4(b). Precision and Recall curve for Descriptive queries.**

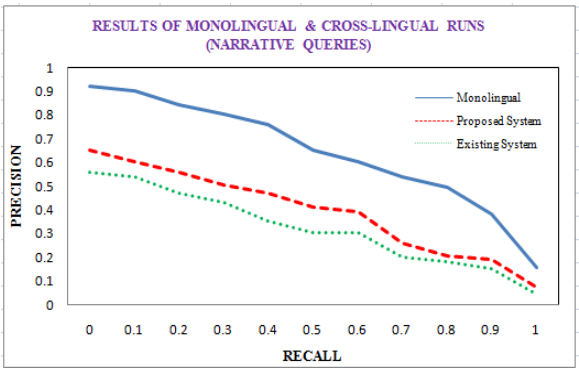


**Fig. 4(c). Precision and Recall curve for Narrative queries.**

## 6. Conclusion

The paper developed a Tamil to English CLTR system which translates Tamil queries into English queries using hybrid machine translation approach and retrieve documents using monolingual search engine. The proposed hybrid MT system is a combination of both rule based approach and statistics approach. Rule

based MT system involves several tasks such as tokenization, pre-processing, and translation. The transliteration is carried out using statistical MT system. The proposed Tamil query based translation system showed effective results when compared to existing system. This work translates only the Tamil queries to English language, and in future it can be extended to analyse the performance of Tamil document to English document translation process and vice versa. Semantic or ontology based text retrieval can also be probed and combined with the proposed classification algorithm in the future.

## References

1. Vanopstal, K.; Stichele, R.V.; Laureys, G.; and Buysschaert, J. (2010). Assessing the impact of English language skills and education level on Pub Med searches by Dutch-speaking users. *Proceedings of the seventh International Conference on Language Resources and Evaluation, European Language Resources Association*, Malta, 2944-2948.
2. Arun, P.D. (2014). Retrieved October 5, 2015, from http://trak.in/tags/business/2014/11/19/india-300m-internet-users-2014
3. Global-reach (2004). Global internet statistics. (By language). Retrieved July 3, 2015, from http://global-reach.biz/globstats/index.php3
4. Ogden, D.; Cowie, J.; Davis, M.; Ludovik, E.; Molina-Salado, H.; and Shin, H. (1999). Getting information from documents you cannot read: An interactive cross-language text retrieval and summarization system. *Proceedings of Joint ACM Digital Library/SIGIR Workshop on Multilingual Information Discovery and Access (MIDAS).*
5. Abdelali, A.; Cowie, J.; Farwell, D.; and Ogden, D. (2004). UCLIR: a multilingual information retrieval tool. *Inteligencia Artificial*, 8(22), 103-110.
6. Iswarya, P.; and Radha, V. (2012). Cross language text retrieval: a review. *International Journal of Engineering Research and Applications*, 2(5), 1036-1043.
7. Lehtokangas, R.; Keskustalo, H.; and Jarvelin, K. (2006). Experiments with dictionary-based CLIR using graded relevance assessments: Improving effectiveness by pseudo-relevance feedback. *Information Retrieval*, 9(4), 421-433.
8. Kumar, A.; and Das, S. (2013). Pre-retrieval based strategies for cross language news story search. *Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation*, ACM, New York, Article 11, 1-10.
9. Saraswathi, S.; Asma Siddhiqaa, M.; Kalaimagal, K.; and Kalaiyarasi, M. (2010). Bilingual information retrieval system for English and Tamil. *Journal of Computing*, 2(4), 85-89.
10. Peters, C.; and Sheridan, P. (2000). *Multilingual Information Access*. European Summer School in Information retrieval, Italy.
11. Oard, D.W.; and Dorr, B.J. (1996). *A survey of multilingual text retrieval*. Technical Report. University of Maryland, College Park, MD, USA.
12. Ture, F.; and Lin, J. (2014). Exploiting representations translation from statistical machine cross-language information retrieval. *ACM Transactions on Information Systems*, 32(4), 1-32.

13. Goto, I.; Utiyama, M.; Sumita, E.; and Kurohashi, S. (2015). Pre-ordering using a target-language parser via cross-language syntactic projection for statistical machine translation. *ACM Trans. Asian Low-Resource Language Information Processing*, 14(3), 1-23.
14. Okpor, M.D. (2014). Machine translation approaches: issues and challenges, *International Journal of Computer Science Issues*, 11(5), 159-165.
15. Euro matrix project. (2007). Retrieved October, 2015, from http://www.euromatrix.net/deliverables/Euromatrix_D1.3_Revised.pdf
16. Seetha, A.; Das, S.; and Kumar, M. (2007). Evaluation of the English-Hindi crosses language information retrieval system based on dictionary based query translation method. *Proceedings of 10th International Conference Information Technology*.
17. Pemawat, V.; Saund, A.; and Agrawal, A. (2010). Hindi - English based cross language information retrieval system for Allahabad Museum. *Proceedings of International Conference on Signal and image processing (ICSIP).*
18. Bandyopadhyay, S.; Mondal, T.; Naskar, K.; Ekbal, A.; Haque, K.; and Godavarthy, S.R. (2007). Bengali, Hindi and Telugu to English Ad-hoc bilingual task at CLEF 2007. *Proceedings of Advances in Multilingual and Multimodal Information Retrieval.*
19. Jagarlamudi, J.; and Kumaran, A. (2007). Cross-lingual information Retrieval system for Indian languages. *Proceedings of 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007).* Budapest. Hungary.
20. Pingali, P.; and Varma, V. (2007). IIIT Hyderabad at CLEF 2007 - Adhoc Indian language CLIR task, *Proceedings of 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007).* Budapest. Hungary.
21. Antony, P.J.; Ajith, V.P.; and Soman, K.P. (2010). Kernel method for English to Kannada transliteration. *Proceedings of Recent Trends in Information, Telecommunication and Computing (ITC)*. Kochi, 336-338.
22. Rao, P.R.K.; and Devi, S.L. (2010). AU-KBC FIRE2010 Submission - Cross Lingual Information Re-trieval Track: Tamil- English. *In Working Notes for the Forum for Information Retrieval Evaluation (FIRE) Workshop.*
23. Chinnakotla, M.K.; Ranadive, S.; Bhattacharya, P.; and Damani, O.P. (2007). *Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007*. Advances in Multilingual and Multimodal Information Retrieval, Springer-Verlag, Berlin, 111-118.
24. Saravanan, K.; Udupa, R.; and Kumaran, A. (2013). *Improving Cross-Language Information Retrieval by Transliteration Mining and Generation*. Multilingual Information Access in South Asian Languages, Springer-Berlin, 310-333.
25. Manikandan, B.; and Shriram, R. (2011). A novel Approach for cross language information retrieval. *Proceedings of Third International Conference on Electronics Computer Technology*. Kanyakumari, 34-38.
26. Shriram, R.; and Sugukumaran, V. (2009). Cross Lingual Information Retrieval Using Data Mining Methods. *Proceedings of the Fifteenth Americas Conference on Information Systems.* San Francisco. California, 1-10.

27. Thenmozhi, D.; and Aravindan, C. (2009). Tamil-English Cross Lingual Information Retrieval System for Agriculture Society. *Proceedings of Tamil internet Conference*. Germany, 173-178.
28. Chaware, S.M.; and Srikantha, R. (2009). Domain Specific Information Retrieval in Multilingual environment. *International journal of recent trends in Engineering and technology*, 2(4), 179-181.
29. Maria Gabriela, F.D. (2005). *A Machine Translation Approach to Cross Language Text Retrieval.*USA: Dissertation.com
30. Lakshmana, P.S.; and Kumaran, K. (2012). Machine Translation from English to Tamil using Hybrid Technique. *International journal of Computer Applications*, 46(16), 36-42.
31. Saraswathi, S.; Siddhiqaa, M.; Kalaimagal, K.; and Kalaiyarasi, M. (2010). Bi-lingual Information Retrieval System for English and Tamil. *Journal of Computing*, 2(4), 85-89.
32. Dhanalakshmi, V.; Anandkumar, M.; Shivapratap, G.; Soman, K.P.; and Rajendran, S. (2009). Tamil POS tagging using linear programming. *International journal of recent trends in engineering*, 1(2), 166-169.
33. Ekbal, A.; and Saha, S. (2013). Simulation Annealing based classifier ensemble techniques: Application to part of speech tagging. *An International Journal on Multi-sensor, Multi-Source, Information fusion*, 14(3), 288-300.
34. Iswarya, P.; and Radha, V. (2015). Improved tagging approach for Part-of-Speech in Tamil Language using an ensemble. *International Journal of Applied Engineering Research*, 10(6), 14015-14028.
35. Anand Kumar, M.; Dhanalakshmi, V.; Soman, K.P.; and Rajendran S. (2010). A sequence labelling approach to morphological analyser for Tamil language. *International Journal on Computer Science and Engineering*, 2(6), 1944-1951.
36. Yarowsky, D. (1995). Unsupervised word sense disambiguation rivalling supervised methods. *In Association of Computational Linguistics (ACL) 33*. Cambridge. MA, 189-196.
37. Adriani, M. (2002). English-Dutch CLOR using query translation techniques. Evaluation of Cross-Language Information Retrieval Systems. *Lecture notes in Computer Science*, 2406, 219-225.
38. Deng, L.; and Li, X. (2013). Machine learning paradigms for speech recognition: An overview, *IEEE Transactions on Audio, Speech and Language Processing*, 21(5), 1-30.