ISSN 2395-8618

Cross-Language Information Retrieval with Incorrect Query Translations

Rajendra Prasath and Sudeshna Sarkar

***Abstract*—In this paper, we present a Cross Language Information Retrieval (CLIR) approach using corpus driven query suggestion. We have used corpus statistics to gather a clue on selecting the right query terms when the translation of a specific query is missing or incorrect. The derived set of queries are ranked to select the top ranked queries. These top ranked queries are further used to perform query formulation. Using the re-formulated weighted query, we perform cross language information retrieval. The results are compared with the results of CLIR system with Google translation of user queries and CLIR with the proposed query suggestion approach. We have English and Tamil corpus of FIRE 2012 dataset and analyzed the effects of the proposed approach. The experimental results show that the proposed approach performs well with the incorrect translation of the queries.**

***Index Terms*—Cross-language information retrieval, incorrect query translations, corpus-driven query suggestion, query representation, retrieval performance.**

I. Introduction

ALL information may not be available in all languages. Suppose a user may query for some information in one language. The information may not be present in that language but may be available in another language that could fulfil their information needs. To support users to access information present in a different language, we require information retrieval systems for different language pairs. Such system are called *Cross Language Information Retrieval* (CLIR) systems. The language of the user query is referred to as *Source Language* (SL) and the language in which information is sought is the *Target Language* (TL).

In the simplest implementation of a CLIR system, a query given in the source language needs to be translated in the target language. For this, one may use a SL-TL bilingual dictionary or any other available SL-TL machine translation system. In Natural Language Processing(NLP), the same concepts may be expressed by different terms or phrases. This is called *Synonymy*. Also a term or a phrase can have multiple meanings. This is referred to as *polysemy*. These variations create problems for monolingual searches, but the effects are more in cross language retrieval. The translated query may not be able to retrieve document in the target language because the concepts may be expressed in the target language using different terms. Secondly the bilingual SL-TL dictionary may be incomplete and the query terms may not have right mapping to a query term in the target language. Even if a dictionary is large, it may be not have coverage for technical terms, named entities and so on. Such terms occur very commonly in a user query. Thirdly the SL-TL translation system may be inaccurate and the terms in the source language may be translated to wrong terms in the target language.

There may occur several issues in the translation process of the query from SL to TL. The translation process may result in the following issues:

1) some query terms may not be translated because they are absent in the dictionary.
2) some query terms may be wrongly translated.

In some cases, even when a query term is translated from the source to the target language, the translated term may not be appropriate.

We have listed three queries in Table 1. In this table, the first column shows the original query in Tamil language, the second column shows the actual query intent of user information needs, and the third column shows the dictionary based translation of these three query in the English.

Let us look at the queries listed in Table 1. While using the dictionary based query translation, the query terms underlined in the first column are not translated from source language to the target language. In the first query, the query terms "vengai" has a correct translation in the dictionary: *leopard*. But this query term has another correction translation: *vengai tree* (a kind of tree known for its strength) which is not found in the dictionary. In the context of the query, the term, *vengai tree* is the right translation. In the second query, the query terms *doosu* (*dust* in English) and *padindha* (*ingrained* in Enlgish) are not found in the dictionary. In the third query, the query term *velli* has three different correct translations: *day in a week* or *moon* or *silver metal*. Out of these three senses, the query term *moon* is the correct translation and its sense is appropriate to the actual context of the query.

There may also exist a case in which we may not be able to find the translation of a compound term in the dictionary. For example, the Google translation tool may not be able to translate a term: *marachchattam* in the second query. This term is a compound word composed of the terms: *maram* (*tree* in English) and *chattam* (this term has two translations in English: *law* and *reaper* or *frame*). In this case, *wooden*

Manuscript received on February 20, 2016, accepted for publication on June 16, 2016, published on October 30, 2016.

Rajendra Prasath is with the Department of Computer and Information Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway (email: drrprasath@gmail.com; see http://www.mike.org.in/rajendra).

Sudeshna Sarkar is with the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, 721 302, India (e-mail: shudeshna@gmail.com).

ISSN 2395-8618

frame or *wooden reaper* is the correct translation that fits with the context of the actual query.

Sometimes, the correct query translation can be obtained by referring to the entire corpus. Since we can only retrieve documents present in the corpus, we may look for a set of query terms which co-occur together in the given corpus.

In this work, we want to use corpus based evidence for translating the query so that the query is appropriate with respect to the corpus. It is not about whether the translation is good or bad, but we are concerned about whether the query retrieves the documents or not. Secondly, the retrieved documents should satisfy the information needs as expressed in the original query. Here we propose a general methodology that makes use of the corpus in order to find the translation of the query terms used for document retrieval task.

In this paper, we have worked on cross language information retrieval with Tamil-English datasets of FIRE corpus[1] and run experiments for *adhoc* news document retrieval. In this work, the query is given in Tamil language which is the source language and documents retrieved are in English which is our target language.

The paper is organized as follows: The next section presents a comprehensive the review of literature related to various strategies in cross lingual information retrieval. Section III presents motivations and objective of this research work. Then we describe the underlying cross lingual information retrieval problem and the issues associated with CLIR systems in Section IV. Then in Section V, we describe our proposed CLIR approach in the context of Indian language pairs. Then we present our experimental results in Section VI. Finally Section VII concludes the paper.

II. Existing Work

We have presented some work related to the query translation issues in cross language information retrieval. Here we describe some dictionary based approaches related to our proposed approach.

A bi-lingual machine readable dictionary or thesaurus based query translation is well studied for different language pairs in cross language information Retrieval [1], [2], [3], [4].

Xi and Cho [5] proposed a method to automatically construct a dictionary based on co-occurrence from English-Chinese parallel corpus for query translation. They used different approaches to calculate the candidate translation equivalent pair correlation degree.

Hull and Grefenstett [6] developed a multilingual IR systems at Xerox which translated French queries and English documents. Their approach works as follows: after morphological analysis, each term is replaced with its inflectional root and the system forms a translated query by looking at each root in the bilingual transfer dictionary. Missing terms are kept unchanged in the translated query. This translated query is then used to perform monolingual document retrieval.

Ballesteros and Croft [2] proposed phrasal translation approach to handle multi-term phrases in cross language information retrieval. In this work, authors focused on the local context analysis to find words and phrases related to each query. They compared this approach with local feedback approach to address the errors associated with the dictionary based translation of words and phrases.

Gelbukh [7] presented a thesaurus-based information retrieval system that enriches the query with the whole set of the equivalent forms. Their approach considers enriching the query only with the selected forms that really appear in the document base and thereby providing a greater flexibility.

Oard and Ertunc [8] proposed a translation based indexing approach in which translation and indexing processes area integrated to improve query time efficiency. This approach uses machine readable bilingual dictionary in which the document's language is the source language and query language is the target language. The idea is to add every possible translation of each document - language term in the index.

Garain *et* al. [9] described an approach to deal the transliteration of out-of-vocabulary (OOV) terms in English into Bengali to improve English-Bengali cross language information retrieval. They used a statistical translation model as a basis for transliteration, and present evaluation results on the FIRE 2011 datasets. Authors used Indri system with #syn operator to handle OOV terms.

Recently, Ali Hosseinzadeh *et* al. [10] presented a set of experiments in which the impact of applying Google and Bing translation systems for query translation across multiple language pairs has been compared for two very different cross language information retrieval tasks.

III. Objectives

Since the document retrieval process depends on the translation of the user query, getting the correct translation of the user query is of great interest. There could be many issues in getting the right translation. Terms present in the dictionary may have multiple meaning and it is essential to identify and choose the right meaning appropriate for the user information needs. Alternatively, query terms in the source language may or may not be present in the dictionary (for example, name of a person or a place). So the actual query terms in source language has to be mapped appropriately to its related query terms in the target language. So in the presence of resources like incomplete dictionary, inaccurate machine translation system, and insufficient tools, we have to identify the appropriate translation for the original user query. Also there might exist multiple translations for a given query. The right translation pertaining to user information needs has to be identified from multiple translation outputs. The underlying corpus evidence may suggest a clue on selecting a suitable query that could eventually perform better document retrieval.

[1]FIRE corpus is available at: http://www.isical.ac.in/~fire

ISSN 2395-8618

Query in Tamil (Transliterated Query in English)	Actual Meaning of the Query	Dictionary Based Translated Query in English
வேங்கை மரங்கள் கடத்தல் (vengai marangal kadaththal)	Smuggling of special type of trees called "Vengai"	leopard tree trees smuggling passing
தூசு படிந்த மரச்சட்டம் (doosu padindha marachchattam)	A wooden frame with dust ingrained on its surface	a wooden frame
வெள்ளி முளைக்கும் நேரம் (velli mulaikkum neram)	The rising time of the Moon	venus star silver the planet time

Fig. 1. List of query terms in Tamil and English with their meaning in the correct query context

In order to do this, we want to use the corpus in order to find the most appropriate query translation that could be used for better document retrieval.

IV. Cross Language Information Retrieval

In this section, we describe the basic working principles of a cross language information retrieval system. Users search for some information in a language of their choice and we call this language the *source* language. The user looks for information present either in the query language or in a different language which we call the *target* language. Some cross language IR systems first perform the translation of the user query given in the source language to the target language. Then using the translated query, the CLIR system performs document retrieval in that target language and translates the retrieved documents in the source language so that the users can get the relevant information in a language that is different from their own.

In CLIR systems, translation and ranking are two major tasks.

A. Translation Task

In CLIR systems, either a query or a document has to be translated from SL to TL. We describe below both these methods:

a) Query Translation: Since a query is very short and contains a few terms, it is convenient to translate it from SL to TL and this task is much easier than translating the whole document. Then the translated query is used for monolingual retrieval in the target language.

b) Document Translation: Often query translation suffers from certain ambiguities in the translation process, and this problem is amplified when queries are short and under-specified. In these queries, the actual context of the user is hard to capture and this results in translation ambiguity. From this perspective, document translation appears to be more capable of producing more precise translation due to richer contexts.

In this work, query translation is much simpler compared with the document translation. So we used query translation to map the user query from source language to the target language and then performed monolingual document retrieval in the target language.

B. Document Ranking

Once documents are retrieved and translated back into the source language, a ranked list has to be presented based on their relevance to the actual user query in the source language. So a good ranking methodology is important in cross language information retrieval.

V. The Proposed CLIR system

We present an approach to improve the cross lingual document retrieval using corpus driven query suggestion (CLIR-CQS) approach. We have approached this problem of improving query translation process in the cross language information retrieval by accumulating the corpus evidence and using such evidences to re-formulate the user query for better information retrieval. Here we assumed that a pair of languages: (SL, TL) is chosen and a *dictionary* D is given for this pair of languages.

A. Identifying Missing / Incorrect Translations

Any query translation system (using either a dictionary based or machine translation based approach) translates the user query given in the source language SL into the target language TL. For every query term, we may either get one more terms correct meaning from the dictionary. Such terms are referred to as *synonyms*. But there are terms that have multiple meanings. Such terms are referred to as *polysemy* terms. These polysemy terms having multiple meaning may result in multiple terms during the translation. Since the dictionary may have limited number of entries, we may have missing or incorrect translation of the user query in language TL. To compensate for the missing translation of query terms, we could use co-occurrence statistics from the entire corpus. But this would take a substantial amount of query processing time in an online system. So we use an initial set of document

ISSN 2395-8618

for this purpose. We explore additional terms for partially translated query using co-occurrence of terms in this initial set of retrieved documents. We present an approach that handles the missing or incorrect translation of the user query and improves the retrieval of information in the target language TL.

Let us look at a case in which we have a partially correct translation of the original query. In this case, some query terms are translated into the target language and some are not. In case of missing translations, we use the co-occurrence statistics of query terms in language SL and their translated terms in language TL to identify the probable terms for missing translations of query terms that could result in better retrieval of cross lingual information retrieval.

B. Corpus Driven Query Suggestion Approach

In this section, we describe the Corpus driven Query Suggestion(CQS) approach for the missing or incorrect translations.

Let q_{TL} be the translation (may be a correct or partially correct or incorrect translation) of q_{SL}. We consider the case in which some query terms are translated into the target language and some are not. In case of missing translations, we use the co-occurrence statistics of query terms in the initial set of retrieved documents in language SL and their translated terms in language TL to identify the probable terms for missing translations of query terms that could result in better retrieval of cross lingual information retrieval. We say that two terms co-occur if any only if they appear in the same text segment. In our experiments, we used paragraphs as the unit of text segment.

a) Initial Retrieval: At first, the user query in language SL is given to the search engine which performs monolingual document retrieval in the source language SL and retrieve top n documents in that language: $C_{SL,Q}$. From this initial set of documents, segment the text into paragraph units. From these text segments, extract the list of terms that co-occur with any of the query terms in the text segments. Let QCO_{SL} be the list of all terms that co-occur with the original query term in the corpus: $C_{SL,Q;n}$.

b) Query Translation: Now using a bi-lingual dictionary, for each of the original query terms, find its translations in the target language TL. Here we may or may not find the translation for all query terms in the target language. We assume that we are able to find the translation for at least one query term. Then using the translated query Q', we perform monolingual document retrieval in the target language TL and retrieve top n documents: $C_{TL,Q'}$. We identify and extract text segments from these n documents in the target language TL based on paragraphs. From these text segments, we extract the list of terms that co-occur with any of the translated query terms. Let QCO_{TL} be the list of all terms that co-occur with the translated query Q' in the corpus: $C_{TL,Q';n}$.

From these two lists: QCO_{SL} and QCO_{TL}, we organize the terms in the source and target language as shown in Figure 2.

In this figure, each node corresponds to a term. More specifically large circled nodes represent the actual query terms in the source language and small circled nodes are the terms that co-occur with the query terms in the source language. Similarly, each big hexagon shaped node represents the translated query term in the target language and each small hexagon shaped node denotes the terms that co-occur with the translated query term in the target language. As shown in Figure 2, Group (A) contains actual query terms and terms that co-occur with these actual query terms, both in source language. Similarly, Group (B) contains the translated query terms and terms that co-occur with these translated query terms, both in the target language.

To understand the proposed methodology, we present an example illustrated in Figure 3 to show the mapping between the co-occurring terms so that the set of probable terms could be selected for incorrect or missing translations.

We find all terms that co-occur with the query terms in SL using the retrieved set of documents in SL. Similarly we find all terms that co-occur with the translated query terms in TL using the retrieved set of documents in TL. Since the number of co-occurring terms may be very large, we can afford to select only a few of them due to the actual query processing time in online. So we propose a method to score the co-occurring terms. Based on the score, we may select the terms which are more important. The method used for scoring is given below:

Weighting of query terms Using the initial set of documents, we compute the weights of the terms that co-occur with the query terms. We consider the initial set of top n documents retrieved for the user query in the source language SL.

Let $Q = q_1, q_2, \cdots, q_p$ be the query terms and CO_q be the set of terms that co-occur in the same text segment as term q.

We get the set of terms, say QCO_q, that co-occur with all query terms Q as follows:

$$QCO_Q = \bigcup_{i=1}^{p} CO_{q_i} \tag{1}$$

Next, we describe the proposed approach for weighting of co-occurring terms in detail. At first, for a given user query, we retrieve top n documents using any standard monolingual IR system. Then for every term in the actual query Q, we obtain the set of terms that co-occur in the same text segment. In order to get a clue on the importance of the co-occurring terms, we compute a weight for each term. In order to compute the weight, we first define the term frequency and inverse document frequency of a co-occurring term as follows:

The term frequency *tf* of each term is defined as the number of times it occurs in n text segments. The *idf* of a term, say q_i is computed as follows:

$$idf(q_i) = \log \frac{N}{df_{q_i}} \tag{2}$$

The weight of each co-occurring term $ct_i \in QCO_Q$, $(1 \leq$

ISSN 2395-8618

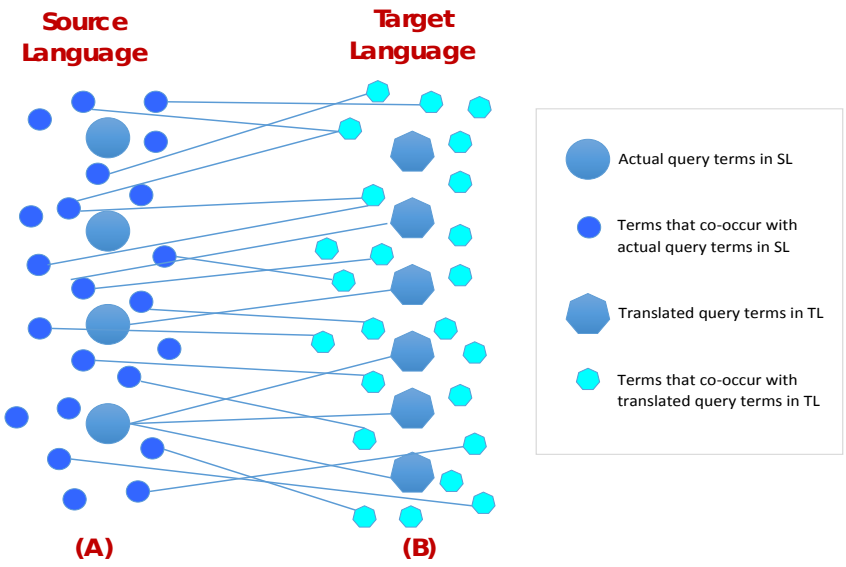

Fig. 2. A conceptual overview of the proposed query suggestion approach

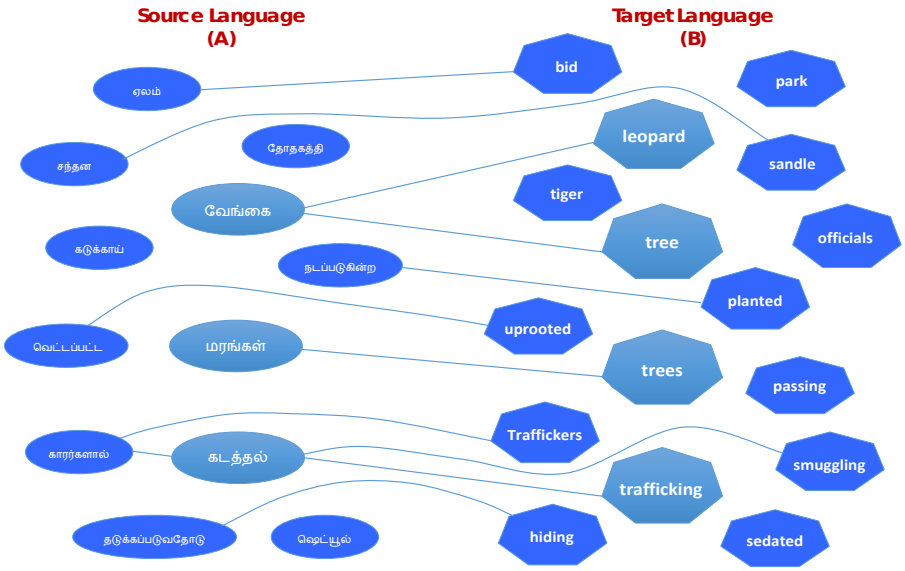

Fig. 3. Example that shows the identification of probable terms for incorrect translation

$i \leq |QCO_Q|$) is computed using the equation:

$$termWeight(ct_i) = tf(ct_i) \times idf(ct_i) \quad (3)$$

Here we explain three different approaches to estimate the weight of the term frequency (in the above equation) across n segments of the retrieved documents:

1) **Term Frequency** (*tf*): The *tf* of a term ct_i is defined as the number of time it occurs in n text segments.
2) **Logarithmic Term Frequency** (log *tf*): Logarithmic value of the term frequency of the term ct_i
3) **Average Term Frequency** (*avg tf*): The *avg tf* of a term ct_i is the ratio between the total number of occurrences of the given term in n text segments and the total number of text segments in which that term occurs.

In this formulation, we use *averageTF* as an indicator for those terms that could either be an entity or the term that tells the type of an entity. For example, named entities may score more term weight giving an indication that its equivalent translation may not exist in the dictionary. Based on the weights, we select the co-occurring terms which are more important in exploring the query terms for missing translations in the target language.

Next, we create a bipartite network by connecting the nodes in the group (A) with the nodes in group (B).

ISSN 2395-8618

c) Bipartite Network: We have two different list of terms: one in the source language SL and another one in the target language TL. Now we create a bipartite network with the terms in QCO_{SL} and QCO_{TL}. Here each term is considered as a node and we add link between a node in QCO_{SL} and a node in QCO_{TL} as follows: A term $q_{SL} \in QCO_{SL}$ is connected to a term $q_{TL} \in QCO_{TL}$ if $< q_{SL}, q_{TL} >$ is found in the dictionary D. We have illustrated an example showing the links between the terms in the SL and the terms in TL in Figure 3.

d) Term Importance: Next we perform the scoring of the co-occurring terms of correct translations in the target language TL. This scoring is used to find a list of candidate terms for missing translations in the target language. We use the bipartite network to find the importance of the terms in QCO_{TL}. For this, we estimate term score $tscore(q_j)$ for each term q_j that has a link to a term in QCO_{SL}.

This term score $tscore(q_j)$ is calculated as follows:

For each term q_j $1 \leq j \leq |QCO_{TL}|$, $tscore(q_j)$ of j^{th} term in QCO_{TL} is computed as:

$$tscore(q_j) = deg(q_j) + \alpha * termWeight(q_j) \qquad (4)$$

where α is a factor to scale the term weights in the retrieved document collection.

e) Identify Probable Query terms for Missing Translations: Based on the computed term scores, we sort the terms in QCO_{TL} and selected the terms with high term scores. Since our method may not be able to get the exact matching terms for missing translations, we add multiple terms that represent different aspects of the missing translation. Then the number of terms with high score is chosen as follows: We assume that a set of topics denoted it by $ntopics$, would be better to represent the missing terms in TL during the translation process. Let nt be the number of terms (in the original query) for which no equivalent translation exists. Now we choose ($ntopics \times nt$) terms from QCO_{TL} associated with each missing query term. This list of probable terms is a representative list for missing or incorrect translation of the query terms in the target language TL.

f) Query Formulation: Using the list of probable terms and their associated term scores, we perform a new weighted query formulation. In this query formulation, we use the term score of each probable term in the target language TL as the boost factor and form a single weighted query. We give more boosting score for the terms for which the one correct translation exists in the dictionary. Otherwise, we distribute the score equally likely to all correct translations. The reformulated weighted query is used by the searcher to perform document retrieval.

g) Document Retrieval and Ranking: Now using the new weighted and reformulated query, we perform document retrieval using BM25 as the ranking function as described in Section VI-A. In fact, we use the default parameters of BM25 ranking approach unchanged.

h) Output: Finally, we return the ranked list of top k documents from the retrieved and ranked set of documents.

We present the pseudocode of the proposed approach in Algorithm V-B0h.

Algorithm 1 CLIR Using Our Query Suggestion Approach

Input: A query having p terms: $Q = \{q_1, q_2, \cdots, q_p\}$ $p > 0$
Index: Documents indexed using Lucene

Description:

1: Get the user query in the source language SL
2: Using this query, retrieve an initial set of n documents in SL
3: Using the dictionary based approach, find the translation of the user query in the target language TL
4: Using this translated query, retrieve an initial set of n documents in the target language
5: Using the documents retrieved for the actual query, identify co-occurring terms of the actual query terms in SL. We call this list as QCO_{SL}.
6: Using the documents retrieved for the translated query, identify the co-occurring terms in the target language TL. We call this list as QCO_{TL}.
7: For each term in QCO_{SL} and QCO_{TL}, we compute a term weight.
8: Based on the term weights, we select the top scoring terms in QCO_{SL} and QCO_{TL} separately.
9: Using the selected top scoring terms, we create a bipartite network: Terms are referred to as nodes. An edge from a node x in QCO_{SL} to a node y in QCO_{TL} is added if and only if the pair $< x, y >$ exists in the dictionary
10: Now compute the term importance score for each term in the target language using the Equation 4
11: Based on the term importance score, rank the terms in QCO_{TL} and choose top d terms with their term importance score.
12: Formulate a single new weighted query in the target language using the terms and their term importance scores which are used as boost factors
13: Perform document retrieval in the target language TL using the newly formed weighted query and BM25 as the ranking function.
14: Generate the final ranked list of documents in the target language TL
15: **return** top k documents in TL as final search results

Output: The ranked list of top k documents in the target language TL.

In the next section, we present the details of our experiments with the proposed cross language document retrieval approach.

VI. Experimental Results

A. Corpus

In this experiment, we considered cross language information retrieval approach on *Tamil* and *English* languages. We

ISSN 2395-8618

have used the multi-lingual adhoc news documents collection of FIRE[2] datasets for our experiments. More specifically, we have used English and Tamil corpus of FIRE 2012 dataset and analyzed the effects of the proposed approach. FIRE 2012 is an incrementally added collection documents from FIRE 2008, FIRE 2010 and FIRE 2011 corpus. The coverage of documents in each of FIRE 2012 collection is listed in Table. I. In this collection, Tamil collection contains news documents more than English collection. English news documents consists of more news documents at the national level where as Tamil news collection covers more regional news.

TABLE I
FIRE 2012 AD HOC DATASET USED IN THIS CLIR EXPERIMENT

Language	# documents	# terms
English	392,577	1,427,986
Tamil	568,335	3,494,299

We have considered a set of 10 queries selected in *Tamil* which are listed in Table II. We have used a Tamil-English bilingual dictionary with 44,000 entries in which there are 20,778 unique entries and 21,135 terms have more than one meaning. We have used this dictionary for translating query terms and also to map the terms co-occurring with the correctly translated pairs.

We use Lucene[3] for indexing and retrieval system with Okapi Best Matching 25 (BM25) ranking used in this paper.

a) Okapi BM25: To rank the final set of the retrieved documents, Okapi BM25 [11], [12] may be used as a ranking function. BM25 retrieval function ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document. Given a query Q and a document D, the similarity score between them is computed using BM25 ranking function as follows:

$$sim(Q,D) = \sum_{i}^{n} idf_i \cdot \frac{tf_{i,D} \cdot (k_1+1)}{tf_{i,D} + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdoclength}})} \quad (5)$$

where $tf_{i,D}$ is the term frequency of the query term i in the document D; $|D|$ is the length of the document D and *avgdoclength* is the average document length in the text collection; k_1, $k_1 \in \{1.2, 2.0\}$ and b, $b = 0.75$ are parameters; and idf_i is the inverse document frequency of the query term i.

Then we retrieve top 20 documents for each query to perform the scoring of candidate terms. We have used three different approaches in weighting the co-occurring terms:

– **Term Frequency**: We use the standard counting of the term frequency across n text segments to compute the weight of each co-occurring term ct_i in SL and TL languages:

$$termWeight(ct_i) = tf(ct_i) \times idf(ct_i) \quad (6)$$

– **Logarithmic Term Frequency**: We use logarithmic function to scale the term frequency using the following equation:

$$termWeight(ct_i) = (1 + \log(tf(ct_i))) \times idf(ct_i) \quad (7)$$

– **Average Term Frequency**: *avg tf* is used to compute the weight of the co-occurring term ct_i in both languages:

$$termWeight(ct_i) = \frac{tf(ct_i)}{m} \times idf(ct_i) \quad (8)$$

where m denotes the total number of segments in which the term ct_i occurs.

We have found that *Average Term Frequency (avgtf)* captures query terms that maps to the translations of the query terms in the target language related to the actual context of the user query. Table III shows the lists of top terms ranked by different term weighting approaches.

Table IV shows the monolingual retrieval performance with FIRE 2012 Tamil Corpus. This is to show the coverage of news documents for the selected query terms in the source language: *Tamil*.

B. Comparisons

We have considered 4 different methods to evaluate the proposed approach.

– **CLIR with Dictionary Based Approach (CLIA-DICT)**: In this experiment, we used the dictionary based approach to translate the user query in source language into the target target language and then documents are retrieved in the target language.
– **CLIR with Google Translation Tool (CLIA-GTT)**: In this experiment, we considered the machine translated query in English using Google translation tool [4] in the period between January 30 and February 9, 2015.
– **CLIR with the Proposed Approach (CLIA-CQS)**: The proposed approach is applied to perform translation of the query from SL to TL using the procedure described in section V-B. In this experiment, we have used top 20 documents are used as the initial set of relevant documents and additional terms are explored for the missing translations.
– **CLIR with the Manual Reference Translation (CLIA-REF)**: Finally we have manually translated the user query into the target language and then performed document retrieval in the target language.

We have manually evaluated the top 10 retrieved documents for each query in the 3-points scale: relevant (1.0), partially relevant (0.5) and irrelevant (0). We used the measure *precision @ top n* documents for each query and tabulated the

[2]Forum for Information Retrieval Evaluation, http://www.isical.ac.in/~fire/
[3]Lucene:www.apache.org/dist/lucene/java/
[4]https://translate.google.com

ISSN 2395-8618

TABLE II
LIST OF QUERIES USED IN OUR EXPERIMENTS

No	Queries in Tamil	Reference Translation in English	Dictionary Based Translation	Translation by Google Translation Tool	Translation by the Proposed approach
1	வேங்கை மரங்கள் கடத்தல்	vengai trees smuggling	leopard tree trees smuggling passing	Leopard trees trafficking	leopard, tree trees smuggling, passing bid scythes tress traffickers planted afforestation axing firewood harms uproot chopping trimming nailed concedes officials
2	தூசு படிந்த மரச்சட்டம்	dirt ingrained wooden frame	a wooden frame	Grime maraccattam	wooden frame clumpy skimmer clods crossovers squalor ingrained railing dislodge mound transformer trays
3	மேற்கில் ஞாயிறு மறைவு	Sun sets in west	the sun as a planet shelter a hiding place secret obscurity	The death Sunday in the West	sun planet secrecy concealment secret hiding place secret obscurity lapses pm skies lifts expiry nightmare disclose
4	சேலம் வீரபாண்டி சிறையில் கலாட்டா	outbreak in Salem Veerapandi prison	in prison comedy	Salem Veerapandi booed in prison	prison comedy hafta slashes coerce panicking sniffs amass prodding extradited accomplice culpable barracks deported
5	சசிகலா ஆதிமுக கட்சியில் இருந்து நீக்கம்	Sasikala expelled from ADMK party	from clearing passing away as clouds darkness fear sleep c, an opening	Shashikala atimuka removal from the party	from clearing passing away as clouds darkness fear sleep c, an opening whines wipeout broadens indistinguishable seeped catcher paradoxically defection disconnect deleted beg boycotting impartial
6	தமிழக மீனவர்கள் போராட்டம்	Tamilnadu fishermen struggle	fishermen diversity of opinions rivalry	Fishermen struggle	fishermen diversity opinions rivalry pirates hostages impounding blockading incarcerated despondent trawlers repatriation encroachment assaults distracted evicted
7	சம்பா பயிர்கள் தண்ணீர் இன்றி வாட்டம்	samba crops fade out without water	cold water distress withered emaciated faintness drooping plants countenance	Samba crops without water gradient	cold water distress, withered emaciated faintness drooping megaliters kuruvai optimized unfeasible eggplant agribusiness percent aquifers contaminating jowar ravaging rice hose overflowed cusecs
8	ஊட்டியில் மலர் கண்காட்சி நிறைவு விழா	closing ceremony of flower exhibition in Ooty	exhibition completion fullness abundance plenteousness completeness festival	Ooty flower show at the closing ceremony	exhibition completion fullness abundance plenteousness, plentifulness, completeness much festival presents pm exhibition paintings tasar workshop armband seamlessly sandalwood art splash
9	கோவையில் முக்கிய பிரமுகர் கைது	important person arrested in Coimbatore	Arrest	The main figure arrested in Coimbatore	arrest gangraped discharged lawful fidayeen offenders escapes conversant arrester tractor assisting disclosure
10	வெள்ளி முளைக்கும் நேரம்	Moon rising time	venus star silver the planet time	Silver germination time	venus star silver planet time weekend flights eclipse astronauts tsunami perigee spaceman amavasya bluish apogee gravitational mavens

results. Table V presents the details of our experiments done in CLIR with Dictionary based Translation (CLIR-DICT); CLIR with machine translation of user queries with Google translation tool (CLIR-GTT); CLIR with the proposed corpus based query selection approach (CLIR-CQS); and CLIR with Manual Reference Translation (CLIR-REF). We used Google translation tool (GTT) [5] to translate the user query given in Tamil language into English language.

C. Discussion

Consider the query ID 1. In this query, there are three tamil query terms: { *Vengai*, *Marangal*, *Kadaththal* }. The term *Vengai* may refer to two variations: *Vengai*, type of a tree whose botanical name is *Pterocarpus marsupium* or *leopard*, and animal; *Marangal*, trees: the correct translation; and finally *Kadaththal* may refer to at least three variations: *trafficking* or *smuggling* or *stealing*. This would give 2 x 1 x 3 = 6 different queries. We identify a set of terms that boosts these query variations and then choose top k terms to form the single weighted query using query terms weighting approach.

[5]Google Translation Tool is available at: http://translate.google.com/

During the evaluation of the proposed approach, we have used 3-points scale for making relevant judgments. We have considered top 10 documents for each query and manually evaluated the retrieved results using the metric: *precision @ top* k documents. The preliminary results show that the proposed approach is better in disambiguating the query intent when query terms that have multiple meanings are given by the users. The average access time for terms set in Tamil is 765.3 milliseconds and 97.8 milliseconds in English. Since the retrieval of initial set of documents and finding co-occurrence terms from this initial set of documents take very negligible amount of time (less than 2 seconds even for top 50 documents), we did not consider the retrieval time comparison in this work.

VII. Conclusion

We have presented a cross language document retrieval approach using corpus driven query suggestion approach. In this work, we have used corpus statistics that could provide a clue on selecting the right query terms when translation of a specific query term is missing or incorrect. Then we rank the set of the derived queries and select the top ranked queries to perform query formulation. Using the re-formulated weighted

ISSN 2395-8618

TABLE III
LIST OF HIGHLY WEIGHTED PROBABLE QUERY TERMS USING THREE DIFFERENT TERM WEIGHTING APPROACHES: tf, log tf AND $avg\ tf$

QID	Actual Queries [Reference Translation]	Highly Weighted Query Terms		
		Term Frequency *(tf)*	**Logarithmic Term Frequency** *(log tf)*	**Average Term Frequency** *(avg tf)*
1	வேங்கை மரங்கள் கடத்தல் [vengai trees smuggling]	2003 reason Twenty year numerous allowed nailed areas oversee curb trees cover	reason Twenty year numerous allowed nailed areas oversee curb trees cover properly decided suffer National fells harms	Smuggling passing bid scythes tress traffickers planted afforestation axing firewood harms uproot chopping trimming nailed concedes officials
4	சேலம் வீரபாண்டி சிறையில் கலாட்டா [outbreak in Salem Veerapandi prison]	2004 1993 hands spoke allowed discussion fighting guards decided Indians Sections framed liable Ensuring represented prison	hands spoke allowed discussion fighting guards decided Indians Sections framed liable Ensuring represented prison landed shortest Qayyum	prison comedy hafta slashes coerce panicking sniffs amass prodding extradited accomplice culpable barracks deported
7	சம்பா பயிர்கள் தண்ணீர் இன்றி வாட்டம் [samba crops fade out without water]	350 2003 318 2015 Water half overflowing year reservoirs areas suffer investing grow dealing Crops impact investment require	Water half overflowing year reservoirs areas suffer investing grow dealing Crops impact investment require easy plant contaminated drums	cold water distress, withered emaciated faintness drooping megaliters kuruvai optimized unfeasible eggplant agribusiness percent aquifers contaminating jowar ravaging rice hose overflowed cusecs
9	கோவையில் முக்கிய பிரமுகர் கைது [important person arrested in Coimbatore]	168 inform thought Nazir Vaiko professional absconding Manoharan farm warrant July blasts based court persons night	inform thought Nazir Vaiko Delhis professional Saturday absconding Manoharan farm warrant July thinks defined blasts based	gangraped discharged lawful fidayeen offenders escapes conversant arrester tractor assisting disclosure

query, cross language information retrieval is performed. We have presented the comparison results of CLIR with Google translation of the user queries and CLIR with the proposed corpus based query suggestion. The preliminary results show that the proposed approach seems to be promising and we are exploring this further with graph based approach that could unfold the hidden relationships between query terms in a given pair of languages.

ACKNOWLEDGMENT

Authors gratefully acknowledge the support extended by Dr. Philip O'Reilly of University College Cork, Cork, Ireland during the last stage of this work.

REFERENCES

[1] G. Salton, "Experiments in multi-lingual information retrieval," Department of Computer Science, Cornell University, Ithaca, NY, USA, Tech. Rep., 1972.

[2] L. Ballesteros and W. B. Croft, "Phrasal translation and query expansion techniques for cross-language information retrieval," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR 1997. New York, NY, USA: ACM, 1997, pp. 84–91. [Online]. Available: http://doi.acm.org/10.1145/258525.258540

[3] J. Capstick, A. K. Diagne, G. Erbach, H. Uszkoreit, A. Leisenberg, and M. Leisenberg, "A system for supporting cross-lingual information retrieval," *Inf. Process. Manage.*, vol. 36, no. 2, pp. 275–289, Jan 2000. [Online]. Available: http://dx.doi.org/10.1016/S0306-4573(99)00058-8

[4] D. Zhou, M. Truran, T. Brailsford, V. Wade, and H. Ashman, "Translation techniques in cross-language information retrieval," *ACM Comput. Surv.*, vol. 45, no. 1, pp. 1:1–1:44, Dec 2012. [Online]. Available: http://doi.acm.org/10.1145/2379776.2379777

[5] S.-M. Xi and Y.-I. Cho, "Study of query translation dictionary automatic construction in cross-language information retrieval," in *Intelligent Autonomous Systems 12*, ser. Advances in Intelligent Systems and Computing, S. Lee, H. Cho, K.-J. Yoon, and J. Lee, Eds. Springer Berlin Heidelberg, 2013, vol. 194, pp. 585–592. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33932-5_54

[6] D. A. Hull and G. Grefenstette, "Querying across languages: A dictionary-based approach to multilingual information retrieval," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR 1996. New York, NY, USA: ACM, 1996, pp. 49–57. [Online]. Available: http://doi.acm.org/10.1145/243199.243212

[7] A. F. Gelbukh, "Lazy query enrichment: A method for indexing large specialized document bases with morphology and concept hierarchy," in *Proceedings of the 11th International Conference on Database and Expert Systems Applications*, ser. DEXA 2000. London, UK, UK: Springer-Verlag, 2000, pp. 526–535. [Online]. Available: http://dl.acm.org/citation.cfm?id=648313.755685

[8] D. W. Oard and F. Ertunc, "Translation-based indexing for cross-language retrieval," in *Advances in Information Retrieval, 24th BCS-IRSG European Colloquium on IR Research Glasgow, UK,*

ISSN 2395-8618

TABLE IV
SELECTED QUERIES IN TAMIL, THE DICTIONARY TRANSLATIONS IN ENGLISH AND THE RETRIEVAL EFFICIENCY IN TAMIL MONOLINGUAL RETRIEVAL

QID	Query in TAmil	Translated Query in ENglish Google Translate / (Derieved Query terms)	User Info Need	p@5	p@10
1	வேங்கை கடத்தல்	Wang conduction / (வேங்கை tree[273] smuggling[110] cut[88] sandle wood[71] tiger[70] வனத்துறையினர்[62] near[50] people[50], steps[45] area[44])	Info about smuggling of Venghai (tree)	0.8	0.65
2	தூசு படிந்த மரச்சட்டம்	Dust-stained maraccattam (dust[128] stained[115] wood[95] coated[75] glass[72] frame[61] time[58] police[52] road[50] people[49] நடவடிக்கை[38])	Info about the dust stained wooden frame	0.7	0.6
3	மேற்கில் ஞாயிறு மறைவு	Sunday on the west side (west[210] india[111] power[106] bengal[105] side[107] sets[101] indies[95] மறைவு[51] ஞாயிறு[48] இரங்கல்[31])	Sun sets on the west	0.6	0.55
4	சேலம் வீரபாண்டி சிறையில் கலாட்டா	Create virapanti Salem in jail (jail[802] வீரபாண்டி[499] ஆறுமுகம்[287] former[149] திமுக[144] court[102] central[98] police[79] authorities[74] prison[70])	Issues made by Salem Veerapandi in prison	0.7	0.5
5	சசிகலா ஆதிமுக கட்சியில் இருந்து நீக்கம்	Athimuka Shashikala from the disposal (சசிகலா[230] அதிமுக[211] party[192] court[166] ஜெயலலிதா[128] disposal[127] chief[118] state[83] minister[82] cases[81])	News about the Sasikala's suspension in ADMK party	0.65	0.6

* Calcutta and Telegraph are the most frequent terms occur in most of the documents. So these terms are not included in our derived query terms

TABLE V
COMPARISON OF RETRIEVAL EFFICIENCY OF TOP 10 SEARCH RESULTS: CLIR-DICT, CLIR-CQS, CLIR-REF AND CLIR-GTT APPROACHES

QID	Precision @ top 5				Precision @ top 10			
	CLIR-DICT	CLIR-GTT	CLIR-CQS	CLIR-REF	CLIR-DICT	CLIR-GTT	CLIR-CQS	CLIR-REF
1	0.05	0.10	0.15	0.20	0.05	0.15	0.25	0.30
2	0.20	0.15	0.40	0.50	0.20	0.25	0.35	0.45
3	0.15	0.10	0.20	0.25	0.15	0.10	0.20	0.25
4	0.15	0.20	0.25	0.30	0.15	0.10	0.20	0.25
5	0.20	0.10	0.35	0.50	0.20	0.20	0.40	0.45
6	0.10	0.05	0.35	0.20	0.10	0.10	0.10	0.15
7	0.05	0.10	0.20	0.30	0.05	0.05	0.20	0.25
8	0.10	0.15	0.10	0.20	0.10	0.10	0.20	0.25
9	0.05	0.10	0.25	0.30	0.05	0.15	0.20	0.25
10	0.10	0.15	0.30	0.40	0.10	0.20	0.20	0.35
Avg	**0.11**	**0.11**	**0.24**	**0.295**	**0.11**	**0.125**	**0.205**	**0.265**

March 25-27, 2002 Proceedings, ser. Lecture Notes in Computer Science, F. Crestani, M. Girolami, and C. J. van Rijsbergen, Eds., vol. 2291. Springer, 2002, pp. 324–333. [Online]. Available: http://dx.doi.org/10.1007/3-540-45886-7_21

[9] U. Garain, A. Das, D. S. Doermann, and D. W. Oard, "Leveraging statistical transliteration for dictionary-based English-Bengali CLIR of OCR'd text," in *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8–15 December 2012, Mumbai, India*, 2012, pp. 339–348. [Online]. Available: http://aclweb.org/anthology/C/C12/C12-2034.pdf

[10] A. Hosseinzadeh Vahid, P. Arora, Q. Liu, and G. J. Jones, "A comparative study of online translation services for cross language information retrieval," in *Proceedings of the 24th International Conference on World Wide Web Companion*, ser. WWW 2015 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015, pp. 859–864. [Online]. Available: http://dx.doi.org/10.1145/2740908.2743008

[11] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proc. of the 17th ACM SIGIR conference on Research and development in IR*, ser. SIGIR 1994. New York, NY, USA: Springer-Verlag New York, Inc., 1994, pp. 232–241. [Online]. Available: http://dl.acm.org/citation.cfm?id=188490.188561

[12] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, Apr. 2009. [Online]. Available: http://dx.doi.org/10.1561/1500000019