Improving Cross Language Information Retrieval Using Corpus Based Query Suggestion Approach

Rajendra Prasath[1,2], Sudeshna Sarkar[1], and Philip O'Reilly[2]

[1] Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur - 721 302, India
{drrprasath,shudeshna}@gmail.com
[2] Dept of Business Information Systems, University College Cork, Cork, Ireland
Philip.OReilly@ucc.ie

Abstract. Users seeking information may not find relevant information pertaining to their information need in a specific language. But information may be available in a language different from their own, but users may not know that language. Thus users may experience difficulty in accessing the information present in different languages. Since the retrieval process depends on the translation of the user query, there are many issues in getting the right translation of the user query. For a pair of languages chosen by a user, resources, like incomplete dictionary, inaccurate machine translation system may exist. These resources may be insufficient to map the query terms in one language to its equivalent terms in another language. Also for a given query, there might exist multiple correct translations. The underlying corpus evidence may suggest a clue to select a probable set of translations that could eventually perform a better information retrieval. In this paper, we present a cross language information retrieval approach to effectively retrieve information present in a language other than the language of the user query using the corpus driven query suggestion approach. The idea is to utilize the corpus based evidence of one language to improve the retrieval and re-ranking of news documents in the another language. We use FIRE corpora - Tamil and English news collections - in our experiments and illustrate the effectiveness of the proposed cross language information retrieval approach.

Keywords: Query Suggestion, Corpus Statistics, Cross-Lingual Document Retrieval, Retrieval Efficiency.

1 Introduction

With the advent of the world wide web, Internet users, speaking a language other than English, are steadily growing. These users create and share information on various topics in their own language and thus the documents in multiple languages grow rapidly over the world wide web. Users cannot access the information written in a language different from their own and hence require a cross language information retrieval(CLIR) system to access information in different languages. In such cross language information retrieval systems, a user

A. Gelbukh (Ed.): CICLing 2015, Part II, LNCS 9042, pp. 448–457, 2015.
DOI: 10.1007/978-3-319-18117-2_33

may query in a source language (known language to the user) and it has to be translated into the target language (unknown language to the user). Then the cross language information retrieval system has to retrieve information, from the unknown language collection, pertaining to the user query in the known language. Since the retrieval process depends on the translation of the user query, getting the correct translation(s) of the user query is of great interest. There could be many issues in getting the right translations.

For a pair of languages chosen by a user, resources, like incomplete dictionary, inaccurate machine translation system, and insufficient tools that could map the term contexts in one language to the similar term contexts in another language, may exist. With these insufficient resources, we have to find a mapping of user queries given in one language to its equivalent query in another language. Also for a given query, there might exist multiple translations. The right translation pertaining to user information needs has to be identified from multiple translations. The underlying corpus evidence may suggest a clue on selecting a suitable query that could eventually perform better document retrieval. To do this, we plan to develop a cross language information retrieval approach based on the corpus driven query suggestion approach. The idea is to use corpus statistics across news documents in different Indian languages and English and then propose a general methodology to utilize the corpus statistics of one language to improve the retrieval of news documents in the other language.

This paper is organized as follows: The next section presents a comprehensive review of literature related to various strategies in cross lingual information retrieval. Section 3 presents motivations and objectives of this research work. Then we describe the underlying cross lingual information retrieval problem and the issues associated with CLIR systems in Section 4. Then in Section 5, we describe our proposed CLIR approach in the context of Indian language pairs. We proceed by presenting our experimental results in Section 6. Finally Section 7 concludes the paper.

2 Existing Work

Capstick *et* al. [1] presented a fully implemented system MULINEX that supports cross-lingual search of the world wide web. This system uses dictionary-based query translation, multilingual document classification and automatic translation of summaries and documents. This system supports *English* and two European languages: *French* and *German*. Gelbukh [2] 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. Zhou *et* al. [3] presented a survey of various translation techniques used in free text cross-language information retrieval. Ballesteros and Croft [4] illustrated the use of pre- and post-translation query expansion via pseudo relevance feedback and reported a significant increase in cross language information retrieval effectiveness over the actual (unexpanded) queries.

McNamee and Mayfield [5] also ensured these findings and showed that pre-translation led to the remarkable increase in retrieval effectiveness where as post-translation expansion was still useful in detecting poor translations. Shin *et* al. [6] presented a query expansion strategy for information retrieval in MEDLINE through automatic relevance feedback. In this approach, greater weights are assigned to the MeSH terms (that are classified for each document into major MeSH terms describing the main topics of the document and minor MeSH terms describing additional details on the topic of the document), with different modulation for major and minor MeSH terms' weights.Levow *et* al. [7] described the key issues in dictionary-based cross language information retrieval and developed unified frameworks, for term selection and the translation of terms, that identify and explain a previously unseen dependence of pre- and post-translation expansion. This process helps to explain the utility of structured query methods for better information retrieval.

3 Objectives

User seeking information may not find relevant information pertaining to his / her information need in a specific language. But information may be available in a different language for his / her information needs, but the user may not know that language. Thus the user may not be able to access the information present in a language that is different from his / her own. To support users to access information present in a different language, cross language document retrieval systems are necessary for different language pairs. In such systems, user query given in a source language has to be translated into the target language and then the cross language retrieval has to be performed.

Since the 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. For a pair of languages chosen by a user, resources, like incomplete dictionary, inaccurate machine translation system, and insufficient tools that could map the term contexts in one language to the similar term contexts in another language, may exist. With these insufficient resources, we have to find a mapping of user queries given in one language to its equivalent query in another language. Also for a given query, there might exist multiple translations. The right translation pertaining to user information needs has to be identified from multiple translations output. The underlying corpus evidence may suggest a clue on selecting a suitable query that could eventually perform better document retrieval. In order to do this, we plan to develop a cross language document retrieval system using a corpus driven query suggestion approach.

4 Cross Language Information Retrieval

In this section, we describe the working principle of a cross language information retrieval system. Users search for some information in a language of their choice

and this language is considered as the source language. Users look for information to be retrieved and presented either in their own choice of the language or in a different language which we consider as the target language. Some cross language IR systems first perform the translation of the user query given in the source language and translates it into the target language. Then using the translated query, the CLIR system performs the 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.

4.1 Issues in CLIR Systems

We list below a few important issues in CLIR systems:

Query Translation: The main issue in CLIR is to develop tools to match terms in different languages that describe the same or similar meaning. In this process, a user is allowed to choose the language of interest and inputs a query to the CLIR system. Then the CLIR system translates this query into the desired language(s).

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 results in translation ambiguity. From this perspective, document translation appears to be more capable of producing more precise translation due to richer contexts.

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 ranking of documents in source and / or target language is essential in cross language information retrieval.

5 The Proposed CLIR System

We present an approach to improve the cross lingual document retrieval using a corpus driven query suggestion (CLIR-CQS) approach. We have approached this problem from enhancing the query translation process in the cross language information retrieval by accumulating the corpus evidence and use the formulate query for better information retrieval. Here we assumed that a pair of languages: (s, t) is chosen and an *incomplete dictionary* (the translation of many terms in the language t may be missing) is given for this pair of languages.

5.1 Identifying Missing / Incorrect Translations

Any query translation system (either based on the dictionary based approach or statistics / example based approach) translates the user query given in the source language s into the target language t. Since the dictionary is incomplete and has limited number of entries, we may have missing or incorrect translation of the

user query in language t. We present an approach that handles the missing or incorrect translation of the user query and to improve the retrieval of information in the target language t.

Let q_i^t be the partially correct translation of q^s. 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 s and their translated terms in language t to identify the probable terms for missing translations of query terms that could result in better retrieval of cross lingual information retrieval (CLIR).

5.2 Corpus Driven Query Suggestion Approach

In this section, we describe the Corpus driven Query Suggestion(CQS) approach for the missing / incorrect translations. Let q_i^t be the translation (may be a correct or partially correct or incorrect translation) of q^s.

In this section, 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 language s and their translated terms in language t to identify the probable terms for missing translations of query terms that could result in better retrieval of cross lingual information retrieval (CLIR). The proposed approach is given in Algorithm. 1.

First we identify the query terms for which the correct translation exists and find the set of co-occurring terms of these query terms. Then we perform weighting of these co-occurring terms. Then we present our procedure to identify the probable terms for missing translations in the actual user query by creating a connected graph using the actual query terms; the co-occurring terms in language and their available translations in the target language.

Weighting of Query Terms: Using corpus statistics, we compute the weight of the terms that co-occur with the query terms as given in Algorithm 2. We consider the initial set of top n documents retrieved for the user query in the source language s.

Scoring Candidate Terms: We perform the scoring of the co-occurring terms of correct translations in the target language as given in Algorithm 3. This generates a list of candidate terms for missing translations in the target language.

5.3 Document Ranking

We have used Okapi BM25 [8,9] as our 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, containing keywords $q_1, q_2, \cdots, q_n$, the BM25 score of a document D is computed as:

$$score(Q,D) = \sum_{i}^{n} idf(q_i) \cdot \frac{tf(q_i,D) \cdot (k_1+1)}{tf(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdoclength}})}$$

Algorithm 1. CLIR using probable terms for missing translations

Require: A machine translation system for query translation

Input: Query q^s having a sequence of keywords in language s

Description:

1. **Initial Set**: Input the user query to the search engine and retrieve the initial set of top n documents in language s: $D = \{d_1^s, d_2^s, \cdots, d_n^s\}$
2. **Co-occurring Terms**: Identify the list of terms that co-occur with each of the query terms. Let these terms list be SET_s
3. **Identify Correct Translation**: Using incomplete dictionary, identify the list of query terms for which the correct translation exists. Then for each correct translation, identify co-occurring terms in the target collection. Let SET_t be the list of these terms.
4. Compute weights of the co-occurring terms in top n documents using corpus statistics using the steps given in Algorithm 2
5. **Identify Probable Terms for Missing Translations of Terms**: For each term in SET_s and SET_t, create a bipartite network using incomplete dictionary as follows: if each term w^s in SET_s has a correct translation w^t in SET_t then draw a link from w^s to w^t. Repeat this for all terms in SET_s.
6. Let PT be the list of probable terms; Initialize the list $PT \leftarrow 0$
7. Compute $tscore(w_p)$ using the procedure given in 3
8. Sort terms in SET_t in decreasing order of $tscore(w_p)$, $1 \leq p \leq |SET_t|$. Choose $l\times$ (# terms for which no translation exists) and add then to the PT, where l denotes the number of aspects a user is interested in.
9. **Query Formulation**: Using the terms in PT, formulate the query by choosing $tscore(w_c)$, $1 \leq c \leq |PT|$ as their weights.
10. **Retrieve**: Now using the formulated query, retrieve the documents in the target collection and sort the documents in decreasing order of their similarity scores.
11. **return** top k documents ($k \leq n$) as the ranked list of documents

Output: The ranked list of top $k \leq n$ documents

Algorithm 2. Weighting of co-occurring terms

Input: SET_s - list of terms that co-occur with each of the query terms

1. Using corpus statistics, compute the weight of each co-occurring terms in top n documents as follows:
2. **for** each co-occurring term ct_j, $(1 \leq j \leq SET_s)$ **do**
3. Compute

$$termWeight(ct_j) = idf(ct_j) \times \frac{\sum_{i=1}^{n} tf(ct_j)}{max_{1 \leq j \leq |SET_s|}(\sum_{i=1}^{n} tf(ct_j))} \quad (1)$$

4. where $idf(ct_j)$ denotes inverse document frequency of the term ct_j.
5. **end for**

Algorithm 3. Scoring candidate terms

Input: SET_s - the list of terms that co-occur with each of the query terms;
SET_t - list of co-occurring terms(of correctly translated terms) in the target language.

1. **for** each term w_p $1 \leq p \leq |SET_t|$ **do**
2. compute $d = \#$terms in SET_s that have outlinks to w_p

$$tscore(w_p) = d + \frac{\sum_{i=1}^{l} termWeight(w_p)}{max_{1 \leq l \leq r}(\sum_{i=1}^{l} termWeight(w_p))} \tag{2}$$

where r denotes the number of terms having inlinks from SET_s

3. **end for**

where $tf(q_i, D)$ is the term frequency of q_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 $idf(q_i)$ is the inverse document frequency of the query term q_i.

The inverse document frequency $idf(q_i)$ is computed as:

$$idf(q_i) = \log \frac{N - df(q_i) + 0.5}{df(q_i) + 0.5}$$

where N is the total number of documents and $df(q_i)$ is the number of documents containing the term q_i.

6 Experimental Results

In this section, we present the experimental results of the proposed cross language information retrieval approach on the selected language pairs: *Tamil* and *English*. We have used the multi-lingual adhoc news documents collection of FIRE[1] datasets for our experiments. More specifically, we have used English and Tamil corpus of FIRE 2011 dataset and analyzed the effects of the proposed approach.

We have considered a set of 10 queries in the language: *Tamil* and for each query in Tamil, we consider the machine translated query in English using Google between the period 30 Jan - 09 Feb 2015 and the manual reference translation in English. The queries are listed in table 1. We have used an incomplete Tamil - English 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. Since we use Lucene [2] as the indexing and retrieval system with BM25 ranking system. Since we retrieve top 20 documents for each query and perform the scoring of candidate terms. The average access time for terms set in Tamil

[1] Forum for Information Retrieval and Evaluation - `http://www.isical.ac.in/~fire/`

[2] `Lucene:www.apache.org/dist/lucene/java/`

Table 1. List of queries

No	Queries in Tamil	Google Translation	Reference Translation
1	வேங்கை மரங்கள் கடத்தல்	Leopard trees trafficking	vengai trees smuggling
2	தூசு படிந்த மரச்சட்டம்	Grime maraccattam	dirt ingrained wooden frame
3	மேற்கில் ஞாயிறு மறைவு	The death Sunday in the West	Sun sets in west
4	சேலம் வீரபாண்டி சிறையில் கலாட்டா	Salem Veerapandi booed in prison	outbreak in Salem Veerapandi prison
5	சசிகலா ஆதிமுக கட்சியில் இருந்து நீக்கம்	Shashikala atimuka removal from the party	Sasikala expelled from ADMK party
6	தமிழக மீனவர்கள் போராட்டம்	Fishermen struggle	Tamilnadu fishermen struggle
7	சம்பா பயிர்கள் தண்ணீர் இன்றி வாட்டம்	Samba crops without water gradient	samba crops fade out without water
8	ஊட்டியில் மலர் கண்காட்சி நிறைவு விழா	Ooty flower show at the closing ceremony	closing ceremony of flower exhibition in Ooty
9	கோவையில் முக்கிய பிரமுகர் கைது	The main figure arrested in Coimbatore	important person arrested in Coimbatore
10	வெள்ளி முளைக்கும் நேரம்	Silver germination time	Moon rising time

Table 2. The selected queries in Tamil; the equivalent 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

is 765.3 milliseconds and 97.8 milliseconds. Since the retrieval of the initial set of documents, and finding co-occurrence terms from this initial set of documents take very neglibile amount of time (less than 2 seconds even for top 50 documents), we did not consider the retrieval time comparison in this work.

Table 3 presents the details of our experiments done in CLIR with machine translation of user queries with Google translation [3] and CLIA with the proposed corpus based query selection approach. We used Google translation to translate the user query given in Tamil language into English language. For every query term, we may either get one or more terms with correct meaning. Now the given

[3] `https://translate.google.com`

Table 3. Comparison of retrieval efficiency of top 10 search results: CLIR with Machine Translation (Google) vs CLIR with the proposed corpus based query suggestion approach

	Precision @ top 5		**Precision @ top 10**	
QID	**CLIR-MT**	**CLIR-CQS**	**CLIR-MT**	**CLIR-CQS**
1	0.10	0.40	0.25	0.40
2	0.15	0.25	0.20	0.45
3	0.10	0.35	0.25	0.35
4	0.20	0.40	0.35	0.55
5	0.10	0.45	0.40	0.50
6	0.15	0.50	0.35	0.60
7	0.10	0.35	0.25	0.40
8	0.20	0.50	0.40	0.55
9	0.10	0.25	0.20	0.35
10	0.15	0.35	0.25	0.45

query terms span over multiple queries with the permutation of the matching query terms (of different meaning). We use corpus statistics to score each of the queries. Then we have considered the top k queries to perform the formulation of a single weighted query. Then using this formulated query, we have performed cross language information retrieval with Tamil-English documents collection. Consider the query ID: 1. In this query, there are 3 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*, *leopard* - animal; *Marangal* - trees - the correct translation; and finally *Kadaththal* - may refer to at least 3 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 the top k terms to form the single weighted query using query terms weighting approach.

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 evaluted 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.

7 Conclusion

We have presented a 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 queries when translation of a specific query term is missing or incorrect. Then we rank the set of the derieved queries and select the top ranked queries to perform query formulation. Using the re-formulated weighted 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 a graph based approach that could unfold the hidden relationships between query terms in a given pair of languages.

References

1. Capstick, J., Diagne, A.K., Erbach, G., Uszkoreit, H., Leisenberg, A., Leisenberg, M.: A system for supporting cross-lingual information retrieval. Inf. Process. Manage. 36(2), 275–289 (2000)
2. Gelbukh, A.: Lazy query enrichment: A method for indexing large specialized document bases with morphology and concept hierarchy. In: Ibrahim, M., Küng, J., Revell, N. (eds.) DEXA 2000. LNCS, vol. 1873, pp. 526–535. Springer, Heidelberg (2000)
3. Zhou, D., Truran, M., Brailsford, T., Wade, V., Ashman, H.: Translation techniques in cross-language information retrieval. ACM Comput. Surv. 45(1), 1:1–1:44 (2012)
4. Ballesteros, L., Croft, W.B.: 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, SIGIR 1997, pp. 84–91. ACM, New York (1997)
5. McNamee, P., Mayfield, J.: Comparing cross-language query expansion techniques by degrading translation resources. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2002, pp. 159–166. ACM, New York (2002)
6. Shin, K., Han, S.-Y., Gelbukh, A., Park, J.: Advanced relevance feedback query expansion strategy for information retrieval in MEDLINE. In: Sanfeliu, A., Martínez Trinidad, J.F., Carrasco Ochoa, J.A. (eds.) CIARP 2004. LNCS, vol. 3287, pp. 425–431. Springer, Heidelberg (2004)
7. Levow, G.A., Oard, D.W., Resnik, P.: Dictionary-based techniques for cross-language information retrieval. Inf. Process. Manage. 41(3), 523–547 (2005)
8. Robertson, S.E., Walker, S.: 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, SIGIR 1994, pp. 232–241. Springer-Verlag New York, Inc., New York (1994)
9. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. 3(4), 333–389 (2009)