Tamil English Cross Lingual Information Retrieval

T. Pattabhi R.K. Rao and Sobha Lalitha Devi

AU-KBC Research Centre, MIT Campus of Anna University
Chrompet, Chennai, India
{sobha@au-kbc.org}

Abstract. This paper describes our work on participation in the FIRE 2010 evaluation campaign in the cross lingual information retrieval track. We describe how cross lingual information retrieval can be effectively performed between a highly agglutinative language, Tamil and English, an isolating language. Agglutination is a morphological process of adding affixes to word base. These affixations can be between noun- noun, adjective-noun, noun-case, etc. This phenomenon of the language has brought serious problems in translation, transliteration and expansion of the query into another language. To overcome these we have used a morphological analyzer which gives the root word or a word base. The word base is used in turn for translation, transliteration and query expansion. The translation of the query is done using bilingual dictionary and transliteration uses statistical method. And query expansion is performed using ontology and WordNet.

Keywords: Cross lingual Information Retrieval (CLIR), Indian languages, Tamil, English.

1 Introduction

The World Wide Web (WWW) or internet today has enormous data in various languages. This is being considered as a huge repository of information, by people all over the world. The cross lingual information retrieval in Indian languages has attracted interest of researchers and industry, only in recent times. One of the first known initiative involving Indian languages was during the TIDES surprise language exercise [9]. In this exercise Hindi, was the surprise language given to the participants. The international evaluation forum CLEF[1] had introduced a special sub-task specific for Indian languages in the year 2007. Here the Indian languages Hindi, Bengali, Marathi and Telugu were considered. In this several approaches such as language modeling coupled with probabilistic transliteration [11], iterative disambiguation [3], using zonal indexing approach [2], using word alignment learned from SMT [5], were used by different participants. The Forum for Information Retrieval Evaluation (FIRE) is an initiative in this direction, for Indian languages.

This paper, describes our participation in the FIRE 2010. Here we participated in the Ad-hoc cross-lingual document retrieval task. The task is to retrieve relevant

[1] http://www.clef-campaign.org/2007.html

P. Majumder et al. (Eds.): FIRE 2010 and 2011, LNCS 7536, pp. 269–279, 2013.

documents in English for a given Indian language query. We have worked on Tamil – English cross lingual information retrieval system. Here the query language is Tamil and the language of the documents to be retrieved is English. In our work, we have focused on query processing, query translation and query expansion. For the query processing we use morphological analyzer. Query translation is a dictionary based approach. During the query translation phase we have focused on proper translation of named entities. For the query expansion, we have used WordNet and the description field of the queries.

The paper is further organized as follows. In section 2 several problems encountered while developing a cross lingual information retrieval system (CLIR) is described. Section 3 describes our approach in solving these problems and how to build the Tamil – English Cross Lingual Information Retrieval. Section 4 discusses the results and finally section 5 gives the conclusion.

2 Issues in CLIR

In a cross lingual information retrieval system, the user gives queries in his/her own language, and the documents to be retrieved are in different language(s). The user query in language L1 is the query language and L2 is the document language. Here we have taken L1 as Tamil and L2 as English. Depending on the nature of the language we have to use the pre-processors. Here Tamil is an agglutinative and inflectional language. The queries in Tamil require to be processed using a morphological analyzer or a stemmer to obtain the base forms of the query terms.

The main issue in any CLIR system is the translation of the query in L1 to L2 and the performance of the system heavily depends on the accuracy of the translation. The translation of queries is not similar to that of document translation though at the outset it seems similar to and simpler than document translation. A query is a short phrase, and not a full sentence, hence language preprocessing such as part-of-speech tagging, chunking are not possible. Considering each word as an independent token, and translating each token into the target language would not be correct in all the cases. A query in most cases is a named entity, or multi-word expression embedded with named entities. In a dictionary based query translation approach, one of the major problems is the coverage of dictionaries. The coverage problem was handled using special dictionaries in the work of Pirkola [12]. Demner-Fushman and Oard [4], in their work have observed that named entities are 50% of the out-of-vocabulary(OOV) words in the query topics. They have also observed that the performance of the retrieval system reduces up to 60% if OOV terms are common in queries and if they are not handled properly. Hence in our present work we have focussed on translations of Named entities. Here we have classified Named entities into three types. Type one which can be transliterated and doesn't require translation. Type two, which requires translation and type three which needs both transliteration and translation. An example for the third type is as follows: in English the named entity “Andhra Pradesh State Road Transport Corporation”, can not be completely transliterated into Tamil. The Tamil equivalent for this is “anthira pradesa manila

pookku varathu kazhagam". This example clearly states the complexity in query translation and shows that in a query there are portions which require translation and transliteration. Another issue for a CLIR system is the ranking of retrieved documents. The objective of ranking is to display the retrieved documents in the order of relevance to the given query.

3 Our Approach

The documents used for Cross lingual retrieval are in English and it consists of 125638 documents (provided by FIRE). We have used Lucene indexer (Lucene[2] is an open source library), which consists of modules for indexing. It is a full-featured text search engine.

The main components in our cross lingual information retrieval system are
i) Language Analyzer
ii) Query Translation engine
iii) Query Expansion
iv) Ranking

3.1 Language Analyzer

The query has to be processed and translated before it is given to the search subsystem. The query language, Tamil, belongs to Dravidian family of languages and it is morphologically rich. It is a verb final language and has a relatively free word order. Its a highly agglutinative language. The words in Tamil are formed by adding suffixes successively to the root word or the base form. Morphophonemic changes occur when the suffixes are added to the root form. The main lexical categories, Nouns and Verbs take inflections. Nouns take number suffixes, case suffixes and postpositions. Nouns have 8 cases viz., Nominative, Accusative, Dative, Locative, Genetive, Instrumental, Sociative and Ablative. Verbs take tense suffix, PNG suffix (Person, Number and Gender agreement) and clitics. In Tamil we can observe a lot of compound nouns. For example the word "ativayirru" which means "abdomen" is combination of two words "ati" (in English this means "below") + "vayirru" (in English this means "stomach"). In the compound words, inflection happens to the last word. In the example stated above the inflection would happen to the last word "vayirru", such as "ativayirril" which means "in the abdomen". Here the locative case suffix "il" is added to the last word "vayirru" [16]. A more detailed description of Tamil morphology and grammar can be found in Lehman [7].

Hence the query requires to be analyzed morphologically to obtain the base form of the word. We have used a Tamil morphological analyzer [17], which is developed using paradigm based approach and uses a finite state Engine (FSA). The system was tested on the corpus obtained from Central Institute of Indian Languages (CIIL),

[2] http://lucene.apache.org/

Mysore, India. This has approximately 3 million words, consisting of several genres such as stories, politics, literature, recipes. The system performs with an accuracy of 97%. For example for the word

(1) " puththakaththil" – puththakam + thth + il
"in the book" – book + oblique stem+ Locative Case

In this example (1), the root word is "puththakam". This has past tense marker "thtth" and locative case marker "il";

The target language English is not morphologically rich and also not an agglutinative language. In English, and many related languages, morphological variation takes place at the right-hand end of a word-form [14]. Hence for this, a simple stemmer can be used. Here we use a stemmer which is an implementation of Porter stemmer algorithm [13]. This algorithm follows suffix stripping based methodology. The algorithm is very simple in concept, with 60 suffixes, two recoding rules and a single type of context-sensitive rule to determine whether a suffix should be removed. Rather than rules based on the number of characters remaining after removal, Porter uses a minimal length based on the number of consonant-vowel-consonant strings (the *measure*) remaining after removal of a suffix.

3.2 Query Translation/ Transliteration

Our approach for query translation is dictionary-based approach. We have made use of a Tamil – English bilingual dictionary, which is of 150K words. As explained in section 2, query translation is one of the most important component of a CLIR system and in our approach we have focused on the proper translation of the Named entities. We have classified Named Entities (NEs) into three types for this purpose. The first type (Type X) is the one which requires only transliteration and no translation. Transliteration is the process of mapping one language word to other language based on the pronunciation. For example the NE, "John", a name of a person written in English, whether in English or Tamil or any other language would be the same and pronounced the same. The same when written in Tamil would be "jaan". The process of transliteration is suitable for Person names, Location names. The second type (Type Y) of NE need to be translated because they have equivalents in the other language. For example the NE "Electricity Board", which is in English, has a Tamil equivalent, "minsaara vaariyam". Such NEs requiring full translation are translated using a bilingual dictionary. The NEs that require translation, instead if they are transliterated then that would lead to most cases no results not being retrieved or in some cases irrelevant results. The third type (Type Z) are the ones which require both transliteration and translation. For example consider the NE in query topic number103, "Bhaglihar hydro-electric power project". In this the word "Bhaglihar" is a place name and requires only transliteration and "hydro-electric power project" requires to be translated. Actually this NE is a case of embedded NE. The NE

identification in the query topics can be done using a automatic NE recognition (NER) engine. But most of the search topics are short and not full sentences, use of a automatic NER engine is practically difficult. The alternative to this is automatic generation of NE lexicon from a huge corpus using a NER engine and using the generated NE lexicon as a look up list during query processing time of CLIR system. We have a NE lexicon for Tamil, which was developed automatically using a huge corpus of data collected from web. A small subset of the data collected from web is manually NE tagged and a NER engine is trained. NER engine uses conditional random fields model of the machine learning techniques. The query translation algorithm is as follows:

i) If the query is a named entity of type X, then transliterate the query using transliteration engine.

ii) Else, if the query is of Type Y then match the whole query with bilingual dictionary entry,

iii) Else, if the query is of type Z then, split the query into two with n-1 terms as one and nth term as one. Now match the n-1 terms and nth term separately with the dictionary entries, if matches substitute, else the same step till all terms are substituted.

iv) Else, if no match found in the dictionary, transliterate those terms using the transliteration engine.

The transliteration engine is a statistical system, which uses a n-grams based approach [1]. This system has been trained using web corpus and tested on the web corpus. This system performs with an accuracy of 81 %. This system produces all possible correct outputs up to a maximum of ten. All these possible transliteration outputs are retained. The name for example written in Tamil as "syamallaa", in English can be represented as, "Shyamala", "Syamala", "Shyamla". This algorithm uses n-gram frequencies of the transliteration units, to find the probabilities. Each transliteration unit is pattern of consonant-vowel (C*V*) in the word.

3.3 Query Expansion

Query expansion, is the process of adding more terms or phrases to the given query. This is done to help the system in retrieving more number of relevant documents. One of the features of natural language is that we have many ways and words to express a same concept or a single object. For example in Tamil "kovil", "koyil", "aalayam" are used synonymously to mean "temple" in English. The other feature is that same word in a language can mean different, in different context. For example "bark" in English can occur as a noun or verb. Query expansion helps in adding more information which would be helpful in obtaining good search results. Query expansion is done i) using synonyms and ii) using the description field of the query document. The synonyms are obtained using WordNet.

English WordNet contains synonyms of words and is a lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory [8]. WordNet is created based on the assumption that there is a mental dictionary in which the words are organised under conceptual fields or semantic domains. In a WordNet, lexical information is organised in terms of word meanings or concepts rather than word forms. We have used two WordNets English and Tamil. The English WordNet is a large lexical database of English, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. (http://wordnet.princeton.edu/). We use the English WordNet 3.0 version available from the Princeton web site.

Tamil WordNet [15] is built in the similar lines of English WordNet. This shows network of semantic relations between Lexical items based on the lexical relations such as synonymy, compatibility, incompatibility (antonymy, etc.), hyponymy, hypernymy, meronymy, holonymy, troponymy, and entailment. This contains major category of words – nouns, verbs, adjectives and adverbs. This consists of total 50497 words and 41013 unique senses. Here we have also made use of the description field of the query document.

3.4 Ranking

Here, we have used the standard Okapi BM25 Model [6]. Given a keyword query Q = {q1, q2,,,qn} and document D, the BM25 score of the document D is as follows:

$$\text{score}(Q,D) = \sum \text{IDF}(qi).\ \frac{(f(qi,D).\ (k1+1))}{f(qi,D)+k1.(1-b+b.(|D|/avgdl))} \quad (1)$$

$$\text{IDF}(qi) = \log .\ \frac{N - n(qi) + 0.5}{n(qi) + 0.5} \quad (2)$$

where f(qi,D) is the term frequency of qi in D, |D| is length of document D, k1 & b are free parameters to be set, avgdl is the average length of document in corpus, N is the total no. of documents in collection, n(qi) is the number of documents containing qi. In our current experiments, we have taken k1 = 1.2 and b = 0.75.

The basic ranking algorithm is customized to suit our needs. Here we introduce a parameter called boost factor in the equation (1), given above. The boost factor is multiplied for each term in the equation (1), before the summation is done, while computing the BM25 score. The original query terms are given a boost factor of 1.5. No boost factor is given to the other new expanded terms in the query. The boost factor of 0.5 times for original query terms is to retain the importance for the user given query terms, than for the query expanded terms.

4 Experiment and Results

The FIRE document collection for Ad-hoc cross-lingual document retrieval task consists of news articles taken from "The Telegraph", which is one of the popular English news magazines available in India. The data from this news magazine is of the time period July 2004 to August 2007. The total number of documents in this collection is 125638 documents. This consists of news articles across the domains such as sports, politics, business, arts, and science. The English documents are indexed using the Lucene indexer. The documents are indexed after stemming and stop word removal. The porter stemmer algorithm is used for stemming the English documents. For ranking we use the slightly modified okapi BM25 algorithm, which includes the boost factors, to obtain better results. In Lucene, the implementation of okapi BM25 ranking function is not provided by default. There are several plugin extensions available which can be used for this purpose. Here we have used the extension plugin provided by Perez-Iglesias [10]. Here we have submitted two runs. The first run is a basic run, where the query expansion module is not implemented. The second run we have the query expansion module implemented. Here we have implemented the Tamil to English cross lingual information retrieval system. The FIRE 2010 topic set consists of 50 topics in Tamil. The queries are generated using the "Title" of the topic set.

We have used standard evaluation measures, which is used in all retrieval tasks. The following evaluation measures are used i) Mean Average Precision (MAP), ii) Precison at 5 (P@5), iii) Precision at 10 (P@10) iv) Precision at 20 (P@20) and v) Recall.

The below table, Table 1, gives the overall results of our submissions in the FIRE.

Table 1. Overall Results of the Tamil – English cross lingual information retrieval

Run ID	MAP	R-Prec	P5	P10	Recall	MAP score as percentage of English Monolingual result in FIRE 2010
2	0.3980	0.3742	0.4640	0.3900	0.9785	77.53%
1	0.2954	0.2931	0.360	0.2960	0.9372	57.54%

On analyzing the results obtained we observe that for queries such as query no. 124, 117, 93, 90 the system did not perform well. The query 124 in Tamil "inthiya maanilangal palavarril cattathirku purrampaka pothaiporull virrpaNai", means "Sale of illegal drugs in various Indian states". This query retrieved all documents consisting the term "drugs", "narcotics" even though those documents do not say about sale of illegal drugs and resulted in retrieval of irrelevant documents. For query 117, the topic is very specific, but the result for this query yields all documents related

to land controversy not just at Kalinganagar. This shows that certain terms in the query should be given negative weightage, so that those terms do not bring in irrelevant documents. Here in this query 117, the terms “land controversy” should be give negative weightage. Similarly for query 93, the system brings in documents describing bribes taken by officials, not just by parliamentarians. This shows the difficulty in tackling specific queries. Handling of specific queries is difficult compared to general queries. In the Table 2, we show all the query topics in English.

Table 2. Query Topic Titles in English

No	Query Title	No	Query Title
76	Clashes between the Gurjars and Meenas	101	Drug party at Pramod Mahajan's bungalow
77	Attacks by Hezbollah guerrillas	102	Pakistani cricketers involved in a doping scandal
78	Conflict between Advani and Singhal over the Ram Mandir issue	103	Bilateral problems surrounding the Baglihar hydro-electric power project
79	Building roads between China and Mount Everest	104	Jaya Bachchan sacked from Rajya Sabha membership
80	Babri Masjid demolition case started against Advani	105	Taj heritage corridor scandal
81	Problems related to the immunization programme against Japanese Encephalitis in India	106	Ban on Taslima Nasreen's novel "Shame"
82	Proposed bus service between Srinagar and Muzaffarabad	107	Furore over the release of a CD containing anti-Muslim sentiments in Uttar Pradesh
83	Election campaign of Laloo Prasad Yadav and Ram Vilas Paswan	108	Greater Nagaland
84	Brinda Karat's allegations against Swami Ramdev	109	New political party formed by Raj Thackeray
85	Abu Salem, accused in the Mumbai Bomb Blast case, in jail custody	110	Sino-Indian relations and border trade
86	Privatization of the Mumbai and Delhi airports	111	Dance bars banned in Mumbai

Table 2. *(Continued)*

87	Discussions between Manmohan Singh and Pervez Musharraf regarding the position of troops around Siachen	112	Links between Gutkha manufacturers and the underworld
88	Popular protests against the arrest of the accused in the Shankar Raman murder case	113	Political clashes in Bangladesh
89	Involvement of Congress ministers in the oil-for-food scam	114	Investigation of the arms scandal in the Defense Ministry
90	Indian representatives visit Bangladesh	115	Serial blasts in Varanasi
91	Allegations of financial corruption against Pratibha Patil	116	Encounter specialist Daya Nayak
92	Activities of the Tamil Tigers of Sri Lanka	117	Controversy over land at Kalinganagar
93	Taking bribes for raising questions in parliament	118	Terrorist strike at Ayodhya
94	Indian Navy accused of leaking classified information	119	Taj Mahal controversy
95	Racism row on the Big Brother show	120	Sex CD scandal involving Anara Gupta
96	Pramod Mahajan's killer	121	Blasts on Samjhauta Express
97	Quarrel between the Ambani brothers regarding ownership of the Reliance Group	122	Sanjay Dutt's surrender
98	India dismisses China's claims on Arunachal Pradesh	123	Death of Yasser Arafat
99	Laloo Prasad Yadav and the fodder scam	124	Sale of illegal drugs in various Indian states
100	Monica Bedi and the passport forgery case	125	Attack on the Lal Masjid

We see that for queries such as 76, 95, 97, 100 etc our system has performed well with MAP scores of 0.800. We observe that the MAP score results for 17 query topics is greater than 0.54 which is comparable with monolingual search result. This was possible because of proper handling of NE terms in the queries. In the query

titles we find that on an average there is at least one NE. Most of the NEs are of type X. Hence the role of transliteration engine is very significant. As explained in section 3.2, when transliterating names from Tamil to English, the correct English spelling form should be produced, else that would lead to irrelevant retrieval results which has happened in the query 77. The term "hespulla" in Tamil was not transliterated properly in English. In the query 78, we had observed that even though the query was translated/transliterated properly the results retrieved was low. The query in Tamil was "athvaani, cinkaal idaiye raamar koyil parriya karuththu veerrupaadu" and in English this was translated as "advani, singhal between ram temple issue conflict". In most of the English documents we found that instead of "temple" they had used "mandir", which is taken from Hindi. This is an interesting characteristic we find in English news articles in India. It would be interesting to study in the corpus, the percentage of such words are in use.

The overall results are encouraging; we obtain a MAP score of 0.3980 when query expansion using synonyms and description field of query is used. This is comparable with English monolingual search. In the FIRE 2010 results we observe that the maximum MAP score obtained for English monolingual search result is 0.5133. Our MAP score is 77.53 % of the monolingual result. From Table 1, we observe that query expansion helps in improving the results significantly. The second implements the query expansion.

5 Conclusion

Here we have presented Tamil to English cross lingual information retrieval system used in the FIRE Ad-hoc evaluation task. Our approach is based on bilingual dictionaries and query expansion. The use of description field of query document gives a significant increase in the recall without disturbing the precision. Here we have found that the system performs well for queries for which the query terms given are unambiguous and world knowledge has been imparted. The overall MAP score of the system is 0.3980 and R-prec is 0.3742. The results are encouraging and comparable to English monolingual system.

References

1. Afraz, M., Sobha, L.: English to Dravidian Language Machine Transliteration: A Statistical Approach Based on N-grams. In: International Seminar on Malayalam and Globalization (2008)
2. Bandyopadhyay, S., Mondal, T., Naskar, S.K., Ekbal, A., Haque, R., Godhavarthy, S.R.: Bengali, Hindi and Telugu to English Ad-hoc Bilingual Task at CLEF 2007. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 88–94. Springer, Heidelberg (2008)
3. Chinnakotla, M.K., Ranadive, S., Damani, O.P., Bhattacharyya, P.: Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 111–118. Springer, Heidelberg (2008)

4. Demner-Fushman, D., Oard, D.W.: The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval. In: 36th Annual Hawaii International Conference on System Sciences (HICSS 2003) – Track 4 (2003)
5. Jagarlamudi, J., Kumaran, A.: Cross-Lingual Information Retrieval System for Indian Languages. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 80–87. Springer, Heidelberg (2008)
6. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments (Part 1 & 2). Information Processing and Management 36(6), 779–840 (2000)
7. Lehmann, T.: A Grammar of Modern Tamil. Pondicherry Institute of Linguistics and Culture, Pondicherry (1989)
8. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An on-line lexical Database (1993)
9. Oard, D.W.: The surprise language exercises. ACM Transactions on Asian Language Information Processing (TALIP) 2(2), 79–84 (2003)
10. Perez-Iglesias, J., Perez-Aguera, J.R., Fresno, V., Feinstein, Y.Z.: Integrating the Probabilistic Models BM25/BM25F into Lucene. CoRR vol. Abs/0911.5046 (2009)
11. Pingali, P., Varma, V.: IIIT Hyderabad at CLEF 2007 Adhoc Indian Language CLIR task. In: Nardi, A., Peters, C. (eds.) Working Notes for CLEF 2007 Workshop (2007)
12. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference, pp. 55–63 (1998)
13. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
14. Sproat, R.: Morphology and Computation. MIT Press, Cambridge (1992)
15. Thiyagarajan, S., Arulmozi, S., Rajendran, S.: Tamil WordNet. In: First Global WordNet Conference, CIIL, Mysore (2002)
16. Vijay Sundar Ram, R., Menaka, S., Sobha, L.D.: Tamil Morphological Analyser. In: Parakh, M. (ed.) Morphological Analysers and Generators, LDC-IL, Mysore, pp. 1–18 (2010)
17. Viswanathan, S., Ramesh Kumar, S., Kumara Shanmugam, B., Arulmozi, S., Vijay Shanker, K.: A Tamil Morphological Analyser. In: Proceedings of International Conference on Natural Language Processing (ICON), Mysore (2003)