Tamil-English Cross Lingual Information Retrieval System for Agriculture Society

D. Thenmozhi and C. Aravindan
Department of Computer Science & Engineering
SSN College of Engineering, Chennai, India
{theni_d, aravindanc}@ssn.edu.in

Abstract

Cross Lingual Information Retrieval (CLIR) system helps the users to pose the query in one language and retrieve the documents in another language. We developed a CLIR system in Agriculture domain for the Farmers of Tamil Nadu which helps them to specify their information need in Tamil and to retrieve the documents in English. In this paper, we address the issue of translating the given query in Tamil to English using Machine Translation approach. It uses a Morphological Analyzer to obtain the root terms of source query. We developed language resources like Bi-lingual Dictionary and Named Entity Recognizer using which the query is translated to English. Local word reordering is performed according to Subject-Verb-Object pattern in order to preserve the relative dependency among the words. Word sense disambiguation is done that identifies the correct sense of an ambiguous word that is being used in a query. The system exhibits a dynamic learning approach wherein any new word that is encountered in the translation process could be updated to the bilingual dictionary. The translated query is given to an existing search engine like Alta Vista, Google, etc. This Machine Translation approach retrieves the pages with Mean Average Precision of 95%. The recall value is also considerably improved.

1. Introduction

The World Wide Web (WWW), a rich source of information is growing at an enormous rate. According to Online Computer Library Center, English is still the dominant language in the web that contributes most of the content [10]. However, global internet usage statistics reveal that the number of non-English internet users is steadily on the rise, but all of them are not able to express their basic needs in English. Tamil users who are not able to express their needs in English are also growing in the Internet. They generally search for the information using the Tamil search engines. But the content provided by these search engines is less in number [13]. Making the huge repository of information on the web, which is available in English, accessible to non-English internet users has become an important challenge in recent times. When the non-English users want to access the existing search engines, most of the time they arrive at improper formulation of English queries.

Cross-Lingual Information Retrieval (CLIR) systems aim to solve the above problem by allowing the users to pose the query in their own (source) language which is different from the language of the documents that are searched. This enables users to express their information need in their native language while the CLIR system takes care of matching it appropriately with the relevant documents in the target language.

CLIR focuses on the cross-language issues from the Information Retrieval perspective rather than Machine Translation perspective [12]. The basic idea in Machine Translation (MT) is to replace each term in the query with an appropriate term or a set of terms from the

lexicon syntactically. If the query is translated based on MT approach, the search will give better result. For example, a Tamil query "*udal nalaththirrku ettra payirkal*" translated to English query "*body health suitable for crops*" in a word by word approach will give an average result. Whereas the machine translation approach translates the query to "*crops suitable for body health*" which gives better result.

We propose a CLIR system using Machine Translation approach in Agricultural domain for Tamil Farmers. The system retrieves relevant documents from an English corpus in response to a query expressed in Tamil language. Here, the query given in Tamil language is translated syntactically and semantically to English (not word by word translation/transliteration) for Information Retrieval process.

Section 2 briefly describes the various works done related to Cross Lingual Information Retrieval systems. Section 3 explain the various phases that are involved in translating the given Tamil query to English using MT approach in Agriculture domain. Section 4 elaborates the various experiments conducted to analyze the performance namely (i) comparison of word by word translation with machine translation, (ii)comparison of Tamil Search Engine with CLIR system and (iii)comparison of irrelevant query formed by non-English users with query translated by CLIR system.

2. Literature Survey

2.1. Cross Lingual Information Retrieval Systems for Indian Languages

Several organizations in India are working on the CLIR system for Indian Languages [13]. Jadavpur University has developed a Bengali, Hindi and Telugu to English CLIR system as part of the ad-hoc bilingual task [15]. IIT Bombay has developed Hindi-English and Marathi-English CLIR systems [10]. IIIT, Hyderabad has developed a Hindi and Telugu to English CLIR system [12]. IIT kharagpur has developed a CLIR system for two most widely spoken Indian languages, Hindi and Bengali [4]. All these works uses bilingual dictionaries. Microsoft Research India has also work on Hindi to English cross-lingual System [8] in which they used a word alignment table that was learnt by a Statistical Machine Translation (SMT) system trained on aligned parallel sentences. These organizations have experimented their results on English corpus of LA Times 2002. AU-KBC had developed Tamil- English Cross Lingual Information Retrieval Track [11] for news articles taken from "The Telegraph", English news magazine in India. All these organizations have developed their CLIR systems using word by word translation approach in news domain.

2.2. Machine Translation Systems

Statistical machine translation (SMT) is an approach to MT that is characterized by the use of machine learning methods. A complete survey and methodologies to build SMT systems is found in [1]. Tamil-English [5] and English-Tamil [2] statistical machine translation system are developed by constructing parallel corpus.

2.3. Applications of Agriculture

Many researches are working on the development of applications in the domain of Agriculture. Food and Agricultural Organization of United Nation has developed an Agricultural Ontology – AGROVAC [7] that provides different concepts and their relations for agricultural domain in different languages of European and Asian languages including Hindi. Knowledge elicitation methods for multiple experts in domain of Agriculture are developed as part of the expert system [3].

We have developed a Cross Lingual Information Retrieval System for Tamil language using MT approach in Agriculture domain.

3. System Architecture

The proposed CLIR system uses a number of phases to translate the given Tamil query in Agriculture domain to an English query using MT approach. This is illustrated in the figure 3.1.

3.1. Morphological Analysis

Morphological Analyzer accepts the input query string and performs a database lookup operation to check whether the given query is directly present in the bilingual dictionary. If present, the translated query is returned. Otherwise split the query into the individual constituent words. By applying morphological rules for handling plurals, case suffices, oblique, etc., the root words are obtained.

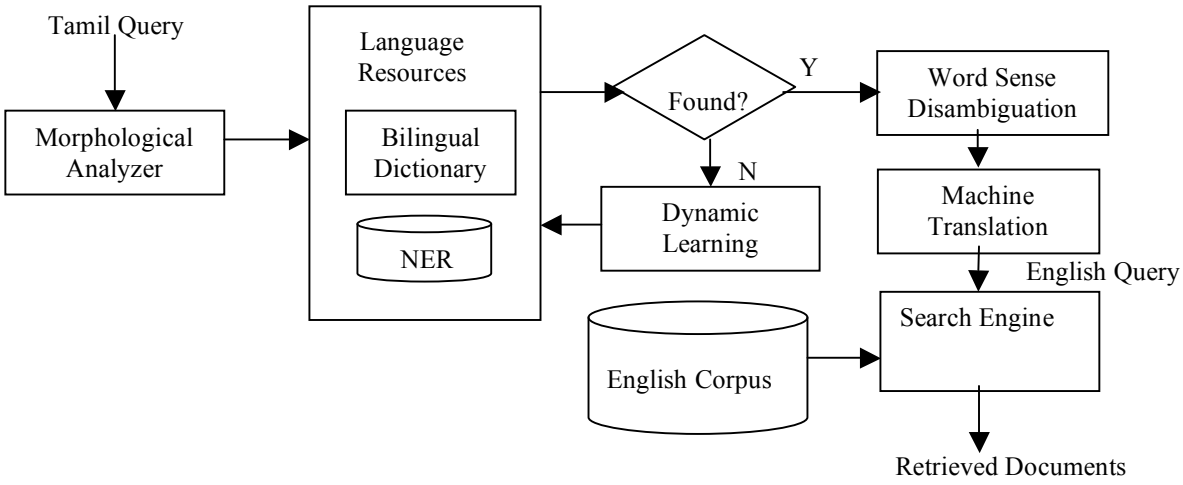

Figure 3.1. System Architecture

3.2. Dictionary Lookup

We have developed a Tamil-English bilingual dictionary of size 5.08MB that contains most the words related to agricultural domain. The dictionary had to be built from the scratch as no resource is available for this domain. After each intermediary step in the Morphological Analyzer, the extracted word is mapped with the bilingual dictionary to check whether it is a root word. If it is available, meaning of the word is returned. If not, the word is then passed on to the subsequent stages in the Morphological Analyzer. At the final stage of the Morphological Analyzer, if the word returned is a root word that is available in the bilingual dictionary, then its meaning in the target language is returned. Otherwise the word is processed so as to bring it to a form that is available in the dictionary and relevant to the context. For example, for the given word "*veelaan*" meaning agriculture, the root word available in the dictionary is "*veelaanmai*". The closest match for "*veelaan*" is identified as "*veelaanmai"* and the meaning is returned.

The system exhibits a dynamic learning approach wherein any new word that is encountered in the translation process could be updated to the bilingual dictionary by

allowing the user dynamically to insert it into the dictionary along with its corresponding English meaning.

3.3. Machine Translation

Tamil is a subject-object-verb (SOV) language. SOV is the type of language in which the subject, object and verb appear in that order. Subject-verb-object (SVO) is a sentence structure where the subject comes first, the verb second and the object third. English is one such language. Tamil to English translation involves classifying the individual translated words into subject, verb and object and placing them in correct ordering. The individual words are processed and identified as to whether they belong to noun or verb and the classification is performed. The words are then arranged according to the SVO pattern to obtain the translated query in English. In order to perform the translation part of speech (pos) tagging should be done for all the words in the dictionary. A local word reordering is performed based on POS tagging to obtain SVO patter of English query [9].

3.4. Word Sense Disambiguation

A complete survey of Word Sense Disambiguation is found in [14]. This phase uses the word-net, a variation of Lesk algorithm [6] to retrieve the possible senses of a word. For each sense of a given word, it is compared with all possible senses of the surrounding words in the given query. The count of number of words common between the sense descriptions is calculated and assigned as the score for the particular sense of the word. The sense that has the highest score is declared the most appropriate one for the target word in the given context. For example, for the query *"aarukalil ulla miin vakaikal"*, the word *"aaru*" is ambiguous having two different meanings, "*The digit six*" and "*River*". The second sense of the word obtains the highest score when compared with the senses of the other words in the query. Thus the correct sense of the word in the given query is "*river*". Hence the query is translated to "*Fish type present in river*".

4. Experimental Results

We have developed a small GUI using which the users can enter their query in Tamil and are translated to English using the CLIR system. The translated query is given to an existing search engines like Alta Vista, Google, etc and the pages are retrieved in English. Various experiments have been done to compare the performance of the developed system with an existing system.

4.1. Precision comparison between Word By Word Translation (WBWT) and MT

To determine the relevance of each retrieved page, a four-point scale was used which enabled us to calculate precision. A page representing full text of research paper, seminar/conference proceedings or a patent is given a score of three and its abstract is given a score of two. A page corresponding to a book or a database is given a score of one. A page representing other than the above (i.e. company web pages, dictionaries, encyclopedia, organization, etc.) is given a score of zero. A page occurring more than once under different URL is assigned a score of zero.

The machine translated queries retrieves documents whose precision is greater than the precision of the documents retrieved using the Word by Word translation Technique which is illustrated in the following table.

Tamil Query	Translated query		Precision(%)	
	WBW trans	Machine Trans	WBWT	MT
Nerppayir saakupatikku ettra urangkal	Paddy crop cultivation for Suitable pesticide	Pesticide suitable for paddy crop cultivation	72	97
Utal nalaththirrku Ettra payirkal	Body health suitable for crops	Crops suitable for body health	69	96
Mann thottarpaana itarpaatukal	Soil related to hurdle	Hurdle related to soil	70	89
Velan thurayil ulla tharpothaya valarssikal	agricultural department present in current development	Current development present in agricultural department	83	91

4.2. Performance comparison between a Tamil search engine and the CLIR System

The non-English(Tamil) users who do not know how to give query in English generally use the Tamil Search Engines. We experimented by giving query in Tamil to Webulagam search and observed that the recall value was very less and the precision was also very low due to the lack of content availability with Tamil Search Engines. We obtained a result with good precision and recall when the same query was given to our CLIR system.

Search System	Webulagam Search	CLIR Search
Query	□□□□□ □□□□□□□□□□	Crop protection
No. of. documents retrieved	57	1,40,000
Precision(%)	23	97

4.3. Search Result and Precision for an Improper query formed by non-English user and Correct query formed using MT

When the non-English(Tamil) users try to formulate their queries in English, most of the time they arrive at improper queries. We have experimented with some improper queries given to an existing search engines and the performance of the search result is low when compared to the query that was translated by the CLIR system.

Search request	Irrelevant query in English	translated to English using CLIR
Munthiri valarkka ettra mann	Cashew grow soil	Soil suitable for cashew growth
No: of documents retreived	14,700	65,700
precision	44	82

5. Conclusion

The CLIR System helps the Farmers of Tamil Nadu, India to pose their information need in Tamil and to retrieve the documents from a large corpus in English language. The system focuses on the Machine Translation technique rather than the word by word translation and gives better result. The CLIR systems generally display the search result in English. It is appropriate, if the results are displayed in their own language for the users who do not know how to give query in English. This system can be further extended to Rank the

pages and provide a summary (in English) of top pages, translate the summary to Tamil or provide an answer to the query in Tamil (like an expert system).

Acknowledgement

We wish to thank C.Karthika and M.Nandhini for their valuable contributions in collecting data related to Agricultural domain, developing the bilingual dictionary and implementing the code for CLIR system. We also thank our management for their continuous motivation and support.

References

1. Adam Lopez, "Statistical Machine Translation**",** ACM Computing Surveys, Vol. 40, No. 3, Article 8, August 2008.
2. Amrita Vishwa Vidyapeetham, "valluvan-English to Tamil Statistical Machine Translation**",** Center for Excellence in Computational Engineering and Networking (CEN), 2005.
3. Bertrand Legar, Oliver Naud, "Experimenting Statecharts for Multiple Experts Knowledge Elicitation in Agriculture, An International Journal on Expert Systems with Apllications, April 2009.
4. Debasis Mandal, Sandipan Dandapat, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar, "Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources", in the working notes of CLEF 2007
5. Fedric C.Gey, "Prospects for Machine Translation of the Tamil Language", in the proceedings of Tamil Internet 2002, California, USA
6. http://en.wikipedia.org/wiki/Lesk_algorithm
7. http://www.fao.org
8. Jagadeesh Jagarlamudi and A Kumaran, "Cross-Lingual Information Retrieval System for Indian Languages", in the working notes of CLEF 2007
9. Maja Popović and Hermann Ney. "POS-based Word Reorderings for Statistical Machine Translation". 5th International Conference on Language Resources and Evaluation (LREC), pages 1278-1283, Genoa, Italy, May 2006
10. Manoj Kumar Chinnakotla, Sagar Ranadive, Pushpak Bhattacharyya and Om P. Damani, "Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007", in the working notes of CLEF 2007
11. Pattabhi R. K. Rao and Sobha L, "AU-KBC FIRE2008 Submission - Cross Lingual Information Retrieval Track: Tamil-English", First Workshop of the Forum for Information Retrieval Evaluation (FIRE), Kolkata. pp 1-5, 2008
12. Prasad Pingali and Vasudeva Varma, "IIIT Hyderabad at CLEF 2007 - Adhoc Indian Language CLIR task", in the working notes of CLEF 2007
13. Prasenjit Majumder, Mandar Mitra Swapan parui and Pushpak Bhattacharyya, "Initiative for Indian Language IR Evaluation", Invited paper in EVIA 2007 Online Proceedings.
14. Roberto Navigli, "Word Sense Disambiguation: A Survey", ACM Computing Surveys, Vol. 41, No. 2, Article 10, February 2009.
15. Sivaji Bandyopathyay, Tapabrata Mondel, Sudip Kumar Naskar, Asif Ekbai, Rejwanuj Haque, Sinivasa Rao Godavarthy, "Bengali, Hindi and Telugu to English ad-hoc Bilingual task at CLEF 2007", in the proceedings of Cross Lingual Evaluation Forum(CLEF) in 2007.