Proceedings of the 2nd National Conference; INDIACom – 2008
Computing For Nation Development, February 08 – 09, 2008
Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi

Dictionary Based Bengali-English Relevant CLIR

Avishek Banerjee, Swapan Bhattacharyya
Department of Computer Science & Engineering and Information Technology,
Asansol Engineering College Sen Raleigh Road, Kanyapur, Asansol, 713304, WB, India.
Email: avishekbenerji@gmail.com, swapanbhattacharyya@hotmail.com

ABSTRACT

CLIR techniques searches unrestricted texts and typically extract term and relationships from bilingual electronic dictionaries or bilingual text collections and use them to translate query and/or document representations into a compatible set of representations with a common feature set. In this paper, we focus on dictionary-based approach by using a bilingual dictionary, with a combination to statistics-based methods to avoid the problem of ambiguity.

KEYWORDS

Query Translation, Disambiguation, Co- Occurrence Tendency, Training Corpora, indexing, ranking, relevancy etc.

INTRODUCTION

This paper describes a system that uses cross-language information retrieval (CLIR) methods to provide search engines with capability of automatic bilingual search. To meet users' needs, there has been intensive research in recent years on Cross-Language Information Retrieval (CLIR), a technique for retrieving documents written in one language using a query written in another language. Bilingual Machine Readable Dictionaries is a good alternative. We have used a Bengali-English pair as an application for the conducted experiments. Here is a big issue for which we have used frequency patterns to count the relevancy avoiding exceptional cases by programming logic.

The whole process can be summarized as:
Feature terms are first extracted from Web documents for each category in the source and target languages. Then, one or more corresponding categories in the other language are determined beforehand by comparing similarities between categories across languages. After that, we use further measures to evaluate ranked retrieval results, thus, developing reliable and informative test collections. We have used frequency patterns to count the relevancy avoiding exceptional cases by programming logic.

Given a variety of keyword occurrences in a document, the goal is to choose fragments, which are:

(i) Maximally informative about the discussion of those terms in the document
(ii) Self-contained enough as to be easy to read, and
(iii) Short enough to fit within the normally strict constraints on the space available for summaries.

The system would be useful for Web users, expanding the international scope of the Web.

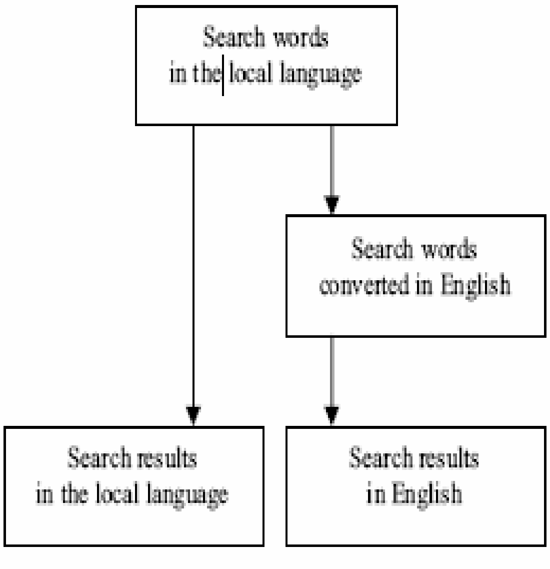

Search words in the local language
Search words converted in English
Search results in the local language
Search results in English

Figure 1. Example of bilingual search

CLIR METHODS

CLIR methods basically includes the following steps
1) Bengali Data dictionary creation.
2) Parsing according to the parts of speech (POS).
3) Creation of stop word list.
4) Bengali Grammar rules Construction.
5) Searching Technique.
6) Indexing and Ranking using relevancy.

1) Bengali Data dictionary creation.

1. Data dictionary is created mainly keeping in views to include the words of Bengali parts of speech (like Nouns, Adjectives, Verb, and Adverb).
2. As these words are the main root words on which our search engine works and they efficiently express the meaning of any sentence or paragraph.

3. These words are included in our data dictionary separately in their combinational form of Unicode Characters (The Unicode Standard 5.0).

4. The present Data Dictionary is created in HTML files as it provide us an easy way to examine the combinational form of the Unicode Characters whether they correctly represent the 'Bangla' words as it is the Standard Dictionary or not.

2) Parsing according to the parts of speech.

The main reasons behind Parsing (Dividing) are:

1) There are various forms of the parts of speech (called "pader rup" in Bengali) in Bengali, which are required to parse so that we can detect the suffixes and/or inflections at the end of the words and add them in the stop word list, which will make our search better.
2) Dividing the search sentences enables a structured way of look since the words are categorized into Nouns, Verbs, Adverbs and Adjectives.

3) Creation of Stop Word List.

অ	অনা	আ	দর	কু	নির	নি	সু
না	পাতি	বি	ভর	ভরা	রাম	স	হা

1. The main purpose of using such stop word list is to separate the non useful words like Conjunctions, Prepositions, Articles, Suffix, and Prefix etc from the query. So that main meaningful Root words can be identified and search can be made using the selected root words.

2. It also includes those "maatraas and bindus" (a Bengali Term) of Bengali Language that bring phonetic changes to the Root words so that the root words can easily be separated and identified.

4) Bengali Grammar rules Construction.

The grammar rules required in preparing and parsing the Data Dictionary are given below. These rules are prepared keeping in mind to differentiate the stop words from the key words. It helps to parse the stop words and remove them from the key words to search from the database where the user in the query gives these words. The grammar rules are as under: -

Rules in Bengali:

Any form of tense will be treated as simple present tense in case of verb. That is when we get the following patterns after the verb they will be treated as same or similar. They are:-

ই	অ	ও	এন	ন	ইস	এ	ইতেছি	ইতেছ
ইতেছেন	ইতেছিস	ই◌েতেছ	ছি	চ্ছি	ছ	চ্ছ	ছেন	চ্ছেন
ছিস	চ্ছিস	ছে	চ্ছে	চ্ছেন	ইয়াছি	ইয়াছ	ইয়াছিস	ইয়াছেন
ইয়াছে	এছি	এছ	এছেন	এছিস	এছে	ইলাম	ইলে	ইলেন
ইলি	ইল	লাম	লে	লেন	লি	ল	লেম	লুম
ইতেছিলাম	ইতেছিলে	ইতেছিলি	ইতেছিল	ইতেছিলেন	ছিলাম	ছিলে	ছিলি	ছিল
ছিলেন	চ্ছিলি	চ্ছিল	চ্ছিলেম	ছিলেম	ছিলুম	ইয়াছিলাম	ইয়াছিলে	ইয়াছিলি
ইয়াছিল	ইয়াছিলেন	ইয়াছিলেম	এছিলাম	এছিলে	এছিলি	এছিল	এছিলেন	এছিলেম
এছিলুম	ইতে	ইতাম	ইতেন	ইত	তাম	তেম	তুম	তে
তেন	ত	তিস	ইব	ইবে	ইবেন	ইবি	ব	বে
বেন	বি	ইতেথাকিব	ইতেথাকিবে	ইতেথাকিবেন	ইতেথাকিবি	তেথাকব	তেথাকবে	তেথাকবেন
তেথাকবি	ইয়াথাকিব	ইয়াথাকিবে	ইয়াথাকিবেন	ইয়াথাকিবি	এথাকব	এথাকবে	এথাকবেন	এথাকবি

Suffix list of verb in Bengali

Prefix list of noun, adjective

প্র	পরা	অপ	সম	নি	অব	অনু	নির	দুর	বি
অধি	সু	উৎ	পরি	প্রতি	অভি	অতি	অপি	উপ	আ

Bengali prefix list (comes from "Sanskrit")

অন্তঃ	আবিঃ	বহিঃ	প্রাদুঃ	তিরঃ	পুনঃ	পূর্ব

Prefixes of some indeclinable-words

ফি	পর	হর	না	ব	বে
বদ	নিম	বর	কার	খাস	খোশ

Bengali prefix list (comes from "Pharsi")

হেড	ফুল	হাফ	সাব

Bengali prefix list (comes from "English")

টি	টা	খানা	খানি	গুলো
গুলি	রাশি	রাজি	বৃন্দ	গণ

সমূহ	আবলী	জাতা	জাতি	

Bengali suffix list of noun (based on “Numbers”).

এর	রা	টার	কে
এঃ	তে	ও	ই
র	এরা		

Bengali suffix list of noun/adjective

এঃ+ রা + ই	এঃ+ রা + ও	এঃ+ র + ই
এঃ+ টার	এঃ+ ও	এঃ+ রা
এঃ+ কে	এঃ+ কে + ও	এঃ+ র + ও
কে + ও	কে + ই	রা + ও

Bengali concatenated suffix list of noun/adjective (from upper list)

1. Steps for Parts Of Speech (POS)
   - i) Go through the search quarry.
   - ii) Decide types the word whether Noun, Verb, Adverb, Adjective or Pronoun from the dictionary format.
   - iii) The Noun, Verb and Adjective will be treated as keywords.
   - iv) Rather than Noun, Adjective, Adverb and Verb all words are treated as Stop Word.
   - v) In Case of verb, noun, adjective there may be a suffix/prefix is attached and that should be identified.

5. Searching Technique

1) The retrieved keywords are matched with the keyword field of database table and corresponding multiple links is retrieved from the database table.

2) The link is referred to the particular HTML file address of the repository and gets the document retrieved from the repository.

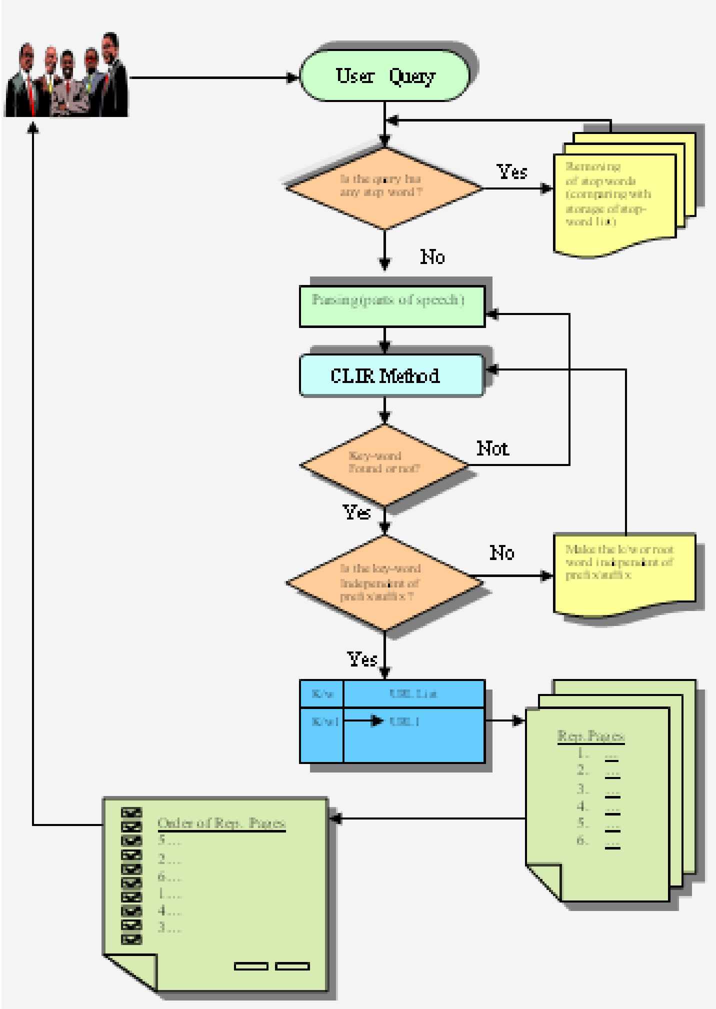
6. Indexing and Ranking using Relevancy

1) The retrieved documents are indexed first with out relevancy checking.

2) The indexed documents are ranked according to their relevancy.

Summary

Here the system first of all identify whether the query includes Bengali terms or not if it contain such words then mainly the Root words are firstly separated form the main query sentence then these separated words are transliterated into equivalent Unicode characters combination of the separated root words, taking these equivalent Unicode combination the equivalent words from the data dictionary is searched if it get any such words from the Data Dictionary it recognize the word and made the search for finding these root words where ever it meet the combination it fetches up and the document from repository is ranked according to the relevancy and displayed to the user in reference link format.


User Query
Is the query has any stop word?
Yes
Removing of stop words (comparing with storage of stop-word list)
No
Parsing (parts of speech)
CLIR Method
Key-word Found or not?
Not
Yes
Is the key-word Independent of prefix/suffix?
No
Make the k/w or root word independent of prefix/suffix
Yes
K/w
URL List
K/w1
URL1
Rep. Pages
1. ...
2. ...
3. ...
4. ...
5. ...
6. ...
Order of Rep. Pages
5...
2...
6...
1...
4...
3...

Results

Search terms: ”kolkata book fair”

Total No of Terms = 4 (“kolkata”, “book”, “fair”,” kolkata book fair”)

CONCLUSION

As Bengali is a well-known and renowned language and Bengali versed population in our country is very high if we can facilitate cross language translation from Bengali to English and vise versa we can achieve to the goal of information searching by religion language and motivate general people toward the use of Information Technology.

FUTURE SCOPE

Relevancy should be a highly sensitive issue in this regard so if we concentrate in this regard and work with large corpora then the system will be checked or tested or verified but for that we need financial help from Government or Corporate House who are interested in this field.

ACKNOWLEDGMENT

This project is going on under the guidance of Dr. P.Mitra, Asst. professor, Department of CSE, IIT, KGP. We are very much thankful for his earnest suggestions; inspiration and involvement that paved the way.

REFERENCES

[1] Stephen Robertson, “TREC Techniques for Cross-language Information Retrieval”, in *Proceeding of the International Workshop on Research Issues in Digital Libraries in co-operation with ACM SIGIR at Vedic Village, Kolkata*, on December 2006.

[2] David A. Hull and Gregory Grefenstette, “Querying across Languages: A Dictionary-based Approach to Multilingual Information Retrieval”, in *Proceeding of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[3] D. A. Hull. Using structured queries for disambiguation in cross-language information retrieval. *Electronic Working Notes of the AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[4] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real user queries on the Web. *Information Processing & Management*, 36(2):207–227, 2000.

[5] F. Kimura, A. Maeda, J. Miyazaki, M. Yoshikawa, and S. Uemura. Cross-Language Information Retrieval Using Web Directory as a Linguistic Resource. *Proceedings of ASI Information Retrieval Symposium (AIRS)*, pages 297–300, 2004.

[6] F. Kimura, A. Maeda, M. Yoshikawa, and S. Uemura. Cross-Language Information Retrieval using Web Directories. *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM ’03)*, pages 911–914, 2003.

[7] C.-J. Lin, W.-C. Lin, G.-W. Bian, and H.-H. Chen. Description of the NTU Japanese-English cross-lingual information retrieval system used for NTCIR workshop. *First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 145–148, 1999.

Ter m(ti)	Frequency (f(ti))	Total term in file	File Containing the term(ti)
Tabular for the 1st file			
Ter m:0	28	34 47	6
Ter m: 1	18		5
Ter m:2	18		5
Ter m:3	8		2
Determined Relevancy----->1.0016707691541733			
Tabular for the 2ndfile			
Ter m:0	67	15 19	6
Ter m: 1	8		5
Ter m:2	9		5
Ter m:3	5		2
Determined Relevancy-----> 1.0021222361091553			