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.

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.

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:
Dictionary Based Bengali-English Relevant CLIR

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

<table>
<thead>
<tr>
<th>এ</th>
<th>রা</th>
<th>টার</th>
<th>কে</th>
</tr>
</thead>
<tbody>
<tr>
<td>এ রা</td>
<td>+ ই</td>
<td>এ</td>
<td>বে ૐ ૌ</td>
</tr>
<tr>
<td>এ</td>
<td>টার</td>
<td>+</td>
<td>ই</td>
</tr>
<tr>
<td>এ</td>
<td>কে</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>কে +</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>

Bengali suffix list of noun/adjective

<table>
<thead>
<tr>
<th>এ</th>
<th>রা</th>
<th>টার</th>
<th>কে</th>
</tr>
</thead>
<tbody>
<tr>
<td>এ</td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>এ</td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>এ</td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

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.

Results
Search terms: “kolkata book fair”
Total No of Terms = 4 (“kolkata”, “book”, “fair”, “kolkata book fair”)

Copy Right © INDIACom – 2008
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