Cross Lingual Information Access System for Indian Languages

CLIA Consortium
A Consortium of 11 Institutions as the Implementing Agency for the
Project "Development of Cross Lingual Information Access (CLIA) System" funded by Government of India, Ministry of Communications & Information Technology, Department of Information Technology (No. 14(5)/2006 – HCC (TDIL) Dated 29-08-2006)

1 Introduction

The **CLIA (Cross Lingual Information Access) Project** is a mission mode project funded by Government of India, Ministry of Communications & Information Technology, Department of Information Technology vide its approval No. 14(5)/2006 – HCC (TDIL), Dated 29-08-2006. It is being executed by a consortium of 11 academic and research institutions and industry partners, IIT Bombay, IIT Kharagpur, IIIT Hyderabad, AU-KBC Chennai, AU-CEG Chennai, ISI Kolkata, Jadavpur University Kolkata, C-DAC Pune, C-DAC Noida, Utkal University Bhubaneswar and STDC, DIT New Delhi. The final deliverables of the project at the end of two years will be a portal where:

- A user will be able to give a query in one Indian language and

- S/he will be able to access documents available in
  (a). the language of the query,
  (b). Hindi (if the query language is not Hindi), and
  (c). English

- Results will be presented to the user in the language of the query. The results can also be presented in the language in which the information originally resided. The languages involved are *Bengali, Hindi, Marathi, Punjabi, Tamil* and *Telugu.*

2 Motivation

With the tremendous growth of digital and online information repositories new opportunities and new problems are created for achieving information retrieval across different languages. Online documents are available internationally in many different languages. Cross Lingual Information Access (CLIA) systems makes it possible for users to directly access sources of information which may be available in languages other than the language of query. However in conventional information retrieval systems the user must enter a search query in the language of the documents in order to retrieve it. This requires that the user can formulate his/her queries in all possible languages and can decipher documents returned by the retrieval process. This restriction clearly limits the amount and type of information, which an individual user really has access to.

Cross-language information retrieval enables users to enter queries in languages they are familiar to, and uses language translation methods to retrieve documents originally created in other languages. Cross-Language Information Access is an extension of the Cross-Language Information Retrieval paradigm. Users who are unfamiliar with the language of documents retrieved are often unable to obtain relevant information from these documents. The objective of Cross-Language Information Access is to introduce additional post retrieval processing to enable users make sense of these retrieved documents. This additional processing may take the form of machine translation of snippets, summarization and subsequent translation of summaries and/or information extraction.

There have been efforts globally towards development of such systems. Cross-Language Evaluation Forum (CLEF), NTCIR Asian Language Retrieval, Question-answering Workshop and others have been working towards achieving the similar goals. In Indian context the need of such system becomes more evident that being multi-lingual country, the people here are familiar with more than one language. The availability of such system

helps in reaching the information if it is available in language other than the language of query. In order to meet this requirement, the CLIA (Cross Lingual Information Access) project has been initiated.

3 System Features

The system is intended to search different documents in Indian languages. Once the user starts the system, an initial screen with logo is displayed. By default, the screen is displayed in Hindi or English depending on the default language selected on the browser of the user. If the user wants to display this initial screen in any other language, he/she can select the language from the bottom of the screen. The screen is then displayed in the selected language. At present, the screen is available in seven languages: Hindi, English, Marathi, Punjabi, Bengali, Tamil and Telugu. To search a document, the first activity the user performs is the selection of the source language. Selection of source language allows the user to enter the text in the selected language.

•Selection of the Source Language: The user can select the source language by clicking a drop-down box. The system displays the languages available to select the source language.

•Entering String for Search: The user enters the query string on which the search is to be made in the appropriate place. The system allows the user to enter the string in the source language selected by the user using a soft keyboard for the language.

•Search the Web or the Site: Once the string is entered, the user should select whether to search the local site or the World Wide Web. The user can then click the search option to search the site for the string entered.

• Displaying the Results: Once the query is properly expanded and translated, it is used to search the web or the local site and the documents are retrieved according to the query. The snippets of the retrieved documents are displayed in the original language of the document as well as in the source language selected by the user. Thus, if the source language selected is Bengali, the user can enter query string in Bengali, the CLIA system searches for documents in English, Hindi and Bengali either from the web or the local site. The snippets of the retrieved documents are displayed in English/ Hindi and Bengali.

•Advanced Search: The user can also select the advanced search option and the CLIA system displays all the options accordingly. The user can select here the domain for which he/she wants to search the documents. At present, the tourism and health domains are available. The user can also select the number of items to be displayed on a single page. By default, the system displays 10 items on a single page. Once the selection is made, the user can click the 'search' option to start the search. In the advanced search option, the CLIA system provides summary as well as extracted information in the form of predefined information extraction templates, of the retrieved documents along with the generated snippet. The summary and the extracted information templates can be displayed in the original language of the document as well as in the source language selected by the user.

4 Technical Details

The CLIA system achieves its performance by means of the following five subsystems:

- Input processing
- Search
- Processing of Retrieved Documents
- Output Generation
- UNL Based Search

The main purpose of each of these subsystems is described below:

- **Input Processing Subsystem**

Input processing analyses the query entered by the user using language processing tools, expands the query to add more relevant terms and based on its analyses, either translates or transliterates all the query terms to the target language and then provides this as input to the search modules. The CLIA Input Processing subsystem consists of Language Analyzer (Tokenization, Named Entity Recognition, Multiword Expression, Stop word identification, Stemmer), Query Expansion (Pre- and Post-Translation Query Expansion), Query Trans-

lation and Query Transliteration. The CLIA Focus Crawler subsystem consists of Classifier, Language identifier and the Parallel crawler.

- **Search Subsystem**

The search subsystem lies at the heart of the CLIA-search engine. The main purpose of this module is to:
(a). Crawl the web and download files for a spe cific language and domain.

(b). Extract the text part in these documents and perform certain processing on those texts and convert them into indices.

(c). Extract results for a particular query by looking up the indices built so far.

(d). Arrange the document references returned by the search subsystem according to some order depending upon page ranking.

- **Document Processing Subsystem**

The document-processing module facilitates the access of documents written in English, Hindi and in the other languages. The documents crawled from the web are preprocessed using language processing tools for extracting information and translating the extracted information into target languages. This module consists of many language-processing tools such as Document Converter, Language Pre-processors, POS taggers, Text Chunker, Named Entity Recognizer, Domain Dictionaries, Information Extraction Engine and Translation engine. These modules are used in processing the documents.

- **Output Generation Subsystem**

This subsystem consists of the snippet generation, summary generation and the snippet translation modules. Brief details of the modules are described below:
(a). Snippet Generation: The Snippet Generation Module generates the snippet corresponding to the retrieved document. This module gets the parse text of the retrieved documents and the query from the search engine and generates the Snippet of each document and returns the generated snippet on the output screen.

(b)**.** Summary Generation: The Summary Generation module generates the summary corresponding to the retrieved document. This module gets the parsed text of the retrieved documents and the query from the search engine and generates the summary of the documents.

(c)**.** Snippet and Summary Translation: Generated snippets for English and Hindi documents are translated to the query language. If the query language is Hindi, then English documents are translated to Hindi. Translated snippet in the query language is displayed on the output screen along with the original snippet.

- **UNL-Based Search Subsystem**

The advanced search system uses UNL as a language independent intermediate representation to enable translation between the languages. The advanced search using UNL is based on concepts, and relations between concepts rather than bag of words. Hence it enables semantic search. Although the current system is designed for Tamil, it can be extended to other languages.

5 Future Roadmap

The functionalities of the CLIA system have been currently developed for Bengali, Marathi and Telugu. The search option has been limited to the crawled documents that are stored and indexed in the CLIA server. The crawled documents are in the tourism domain. At present, the user can provide 3-4 word queries to the CLIA system using soft keyboards for the respective language. The output of the system shows only the snippets in the original language of the document.

The CLIA system is being enhanced to provide full functionalities in the other Indian languages, i.e., Hindi, Tamil and Punjabi. The search option is expanded to provide search facility on the web also. Work is also going on for providing CLIA functionalities in the health domain. In future, snippet translation, summary generation and translation as well as information extraction templates generation and translation are going to be included in the CLIA system. The evaluation engine will judge the CLIA system based on the ranks of the relevant documents in the list of documents retrieved by the system.