# Initiative for Indian Language IR Evaluation

Prasenjit Majumder Mandar Mitra Swapan Kumar Parui
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute, Kolkata
India.
{prasenjit_t,mandar,swapan}@isical.ac.in

Pushpak Bhattacharyya
Dept. of Computer Science and Engineering
Indian Institute of Technology, Bombay
India.
pb@cse.iitb.ac.in

## Abstract

*The Indian subcontinent can be regarded as another Europe, due to its lingual diversity. Geographically, the Indian subcontinent consists of six countries, namely Pakistan, Bangladesh, Nepal, Sri Lanka, Bhutan and India. The total population in this part of the world is about 1,300 million and about 25 official languages are used by this population. Among the major languages of this region, Hindi and Bengali rank among the top ten most-spoken languages of the world. Over the past few years (2000–2007), a large volume of Indian language (IL) electronic documents have come into existence at a growth rate of 700.0 %. The need for developing IR systems to deal with this growing repository is, therefore, unquestionable. Considering this need, the Government of India has recently formed a national consortium of academic and research organizations, that has been entrusted with the task of developing a Cross Lingual Information Access (CLIA) system for Indian language content. This paper will outline the issues that will need to be addressed, and the activities of the newly formed consortium.*

***Keywords:*** *Indian language, Evaluation, Information Retrieval.*

## 1 Introduction

India is a multilingual and multi-script country. The Indian constitution recognizes 23 official languages. Among these languages, Hindi (about 366 million native speakers) and Bengali (207 million native speakers) rank among the top ten most-widely spoken world languages. Hindi is the national language of India; Bengali (also referred to as Bangla), the second most-spoken language in India, is also the national language of Bangladesh. Telugu (66 million speakers), Marathi (63 million speakers), and Tamil (53 million speakers) rank next in terms of number of speakers.

### 1.1 Linguistic Background

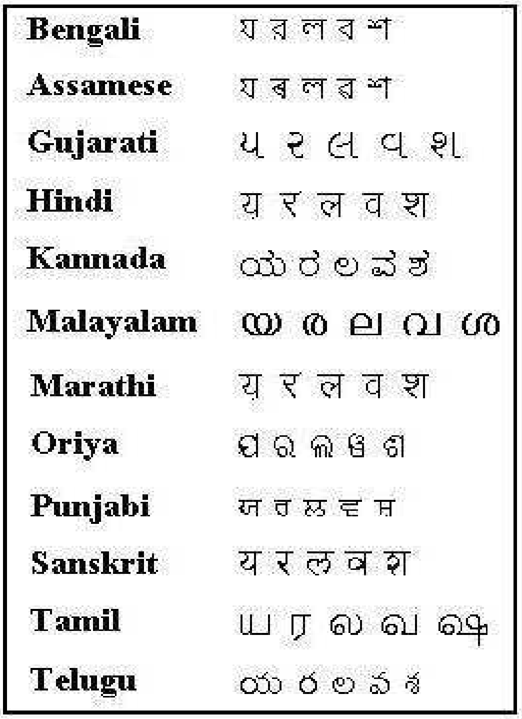
Indian languages belong to the Indo-European class of languages. According to SIL Ethnologue[1], 12 out of the top 20 contemporary world languages (in terms of number of speakers) belong to the Indo-European group. These are Spanish, English, Hindi, Portuguese, Bengali, Russian, German, Marathi, French, Italian, Punjabi and Urdu. Together, they account for over 1,600 million native speakers. Interestingly, 5 out of these are Indian languages. Indian languages have their roots in Sanskrit or the Dravidian languages. These languages share many common grammatical characteristics. The north Indian languages have descended from Sanskrit and the South Indian languages are influenced by Tamil.

### 1.2 Scripts

Writing systems in Indian languages are very old. Most of the languages have their own script. Major scripts in use include Devanagari, Tamil, Bengali, Gujarati, Gurmukhi (Punjabi), Oriya, Telugu, Kannada, Malayalam, Urdu, Marathi, Sindhi and Sinhala. Figure 1 shows a glimpse of various Indian scripts.

With the increasing availability of computers within the Indian sub-continent, the amount of information in electronic form has increased enormously. However, IL electronic documents are stored in several diverse

[1] http://www.ethnologue.com/



| | |
|---|---|
| Bengali | য র ল ব শ |
| Assamese | য ৰ ল ৱ শ |
| Gujarati | ય ર લ વ શ |
| Hindi | य र ल व श |
| Kannada | ಯ ರ ಲ ವ ಶ |
| Malayalam | യ ര ല വ ശ |
| Marathi | य र ल व श |
| Oriya | ଯ ର ଲ ଵ ଶ |
| Punjabi | ਯ ਰ ਲ ਵ ਸ਼ |
| Sanskrit | य र ल व श |
| Tamil | ய ர ல வ ஷ |
| Telugu | య ర ల వ శ |

**Figure 1. Some major Indian Language scripts**

formats and encodings, including proprietary font or glyph-based encodings, the Indian Standard Code for Information Interchange (ISCII), and only recently, Unicode. Documents in ISCII and Unicode encoding can be handled easily, but the real challenge lies in taking care of documents that use proprietary font encodings.

## 2 ILIR Evaluation

Indian Language Information Retrieval is a largely virgin area for IR and NLP researchers. While several groups in India and elsewhere are actively working in this area, it has not yet become a common practice to build basic resources that are made publicly available (or shared within the community). Thus, a pool of shared essential resources — like standard relevance-judged corpora, stemmers, named entity identifiers, multilingual dictionaries, part-of-speech taggers, etc. — is yet to be developed for these languages. More details about available resources can be found in [1].

Among the major events focused on ILIR, the most prominent was the surprise language exercise (SLE) of the DARPA-sponsored TIDES (Translingual Information Detection, Extraction, and Summarization) program[2], where many interesting experiments were carried out for the first time on Indian languages. In this surprise language exercise, the participants started with zero language resources, and constructed workable systems. Although some evaluation resources were quickly developed during that phase, these are inadequate for the sort of detailed experimentation that has been done for resource-rich languages.

## 3 CLIA Consortium

The Government of India has recently launched a two-year, “mission-mode” project on Cross-Lingual Information Access System development for Indian languages. This project is being executed by a consortium comprising ten research groups across the country. The final deliverable 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 access documents available in the language of the query, as well as Hindi (if the query language is not Hindi) and English. In addition to displaying the original documents, the system will also show the search results to the user in the language of the query. The languages involved will be Bengali, Hindi, Marathi, Punjabi, Tamil and Telugu. Table 1 lists the ten participating organizations, along with the language that each is primarily responsible for.

| Organization | Languages |
|---|---|
| AU-CEG Chennai | Tamil |
| AU-KBC Chennai | Tamil |
| CDAC Noida [CDACN] | Hindi, Punjabi |
| CDAC Pune [CDACP] | Bengali, Marathi, Tamil |
| IIIT Hyderabad [IIITH] | Hindi, Telugu |
| IIT Bombay [IITB] | Hindi, Marathi |
| IIT Kharagpur [IITKGP] | Bengali |
| ISI Kolkata [ISI] | Bengali |
| Jadavpur University [JU] | Bengali |
| Utkal University [UU] | |

**Table 1. Participating Organizations in the CLIA Consortium**

In addition to actual system development, the project also includes an evaluation component, to be co-ordinated by Indian Statistical Institute. The evaluation component will basically consist of running TREC-style monolingual and cross-lingual ad-hoc retrieval tasks, as well as a named entity identification task for all these languages. As a fallout, the following benchmark datasets are expected to be created during the course of the project:

a) corpora containing 50,000 documents (or more) for each of Bengali, Hindi, Tamil, Telugu, Marathi, and Punjabi;

b) 50 topics (test queries) and corresponding relevance judgments for each language;

[2] http://www.darpa.mil/ipto/programs/tides/

c) a set of 50 documents from each corpora, manually tagged with named entities.

The corpora will consist of documents crawled from the Internet, and are expected to consist principally of roughly contemporaneous news articles. Besides, domain-specific (sub-)corpora will also be constructed in the areas of health and tourism. The consortium is presently negotiating with publishing agencies for resolving the copyright issues. Following the TREC / CLEF / NTCIR model, each topic (search query) will be divided into title, description and narrative sections.

In order to reduce the amount of work involved in creating the relevance judgments, a single set of topics will be translated into each of the six languages. Thus, the judgments for the monolingual task will also be applicable to the cross-lingual task.

The resources created within the project will be made available to all participants. Additionally, these datasets may be made available to the community of researchers and scientists working on ILIR worldwide. Thus, we hope to have international participation in the first ILIR Evaluation Workshop that is likely to be held some time in 2009.

## References

[1] P. Majumder, M. Mitra, and K. Datta. Multilingual Information Access: an Indian Language Perspective. In *Proc. ACM SIGIR Workshop on New Directions in Multilingual Information Access*, Seattle, 2006.