LANGUAGE IN INDIA
Strength for Today and Bright Hope for Tomorrow
Volume 12 : 8 August 2012
ISSN 1930-2940

Application of NLP in Indian Languages in Information Retrieval

B. A. Sharada, Ph.D.

==

1. Introduction

Information is the need in today's world and is available in abundance in a variety of modes and forms. In order to retrieve the same, research in the interdisciplinary topics have found out many schemes. One such scheme is the application of Natural Language Processing (NLP) that aids to bring humans and the digital world closer. NLP and information retrieval are very different areas of research, and recent major conferences only have a small number of papers investigating the use of NLP techniques for information retrieval (Thorsten Brants). The very fact about NLP is that it facilitates user to search in the original language. In other words we train the computer to learn and retrieve in the natural language using Artificial Intelligence (AI). Now, major of the Indian Languages are on the NLP platform, wherein new technologies are being developed. The present study tries to analyse the research contributions in this field.

1.1 Scope of the Study

The study is limited to Indian Languages and for the years from 2000 to 2010. Such studies globally are more but the same in Indian languages are on the developing side. This is because of the inbuilt complexities of the properties of Indian languages. India, is rich with 22 Scheduled and 100 Non-scheduled Languages spoken by more than 10,000 persons each according to the latest Census 2001.

1.2 Sample Data

Meta data for the research done on the present topic was collected from many sources for the said period such as internet, conference proceedings, reference lists appended to each article, Journal database etc., and while doing the search, each language was searched individually in the internet for its application of NLP.

2 NLP in Indian Languages

Indian languages are not fast in absorbing the technology. So far mainly NLP work is undertaken in 22 Scheduled languages and others remain untouched and even in case of these Scheduled languages, absorption of language technology is not uniform or on the same speed. The NLP work first gets into Hindi and then to other languages.

2.1 NLP and Information Retrieval System

In order to bridge the gap between NLP and information retrieval system (IRS) several tools have been developed. The NLP tools (NLPT) are developed keeping information processing and retrieval also as one of the major objectives. For example, Indian language software tools and fonts for eleven Indian languages mentioned in Table 1 have been developed by C-DAC and released by the Ministry of Communications and Information Technology, Government of India and distributed free of cost and also available at www.ildc.in or www.ildc.gov.in for download.

Abbreviations used in the Table 1:

TTF	True Type Font
KBD	Keyboard Driver
MKE	Multifont Keyboard Engine
UKD	Unicode Compliant Keyboard Driver
GFC	Generic Font Code and Storage Code Converter
SC	Spell Checker
BD	Bilingual Dictionary
DFD	Decorative Font Design Tool
TT	Transliteration Tool
DST	Database Sorting Tool
MSW	Microsoft Word Tool
MSE	Microsoft Excel Tool
TA	Type Assistant
CMS	Content Management System
LTT	Language Typing Tutor
OCR	Optical Character Recognizer
TSS	Text to Speech system

TABLE 1Features in the Language Software Tools

Language	TTF& KBD	MKE	UKD	GFC	SC	BD	DFD	TT	DST	MSW	MSE	TA	CMS	LTT	OCR	TSS
Assamese	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N		
Gujarathi	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N		
Hindi	Y	Y	Y	Y	Y	Y		Y		Y	Y			Y	Y	Y
Kannada	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y
Malayalam	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y		Y	Y	Y
Marathi	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	

Oriya	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y
Punjabi	Y	Y	Y	Y		Y	Y						Y		Y	
Tamil	Y	Y	Y		Y	Y									Y	
Telugu	Y	Y	Y	Y	Y	Y	Y			Y	Y				Y	
Urdu	Y							Y								

Some additional features in the following languages are:

- Assamese - Calendar and Language Scribus for Linux
- Gujarati - Language Tux Paint, Calendar and Language Scribus for Linux, Games and Puzzles
- Hindi - Firefox browser, GAIM Multiprotocol messenger, Email
- Kannada - Library Management System, Text Editor (Nudi), Logo, Personal utilities, Games and Puzzles, Email
- Malayalam - Text Editor, Malayalam/English tutor, Database sorting tool, number to word tool
- Marathi - Marathi WordNet
- Punjabi - Hindi TTF, Hindi MKE, Hindi Unicode compliant, Hindi UKD, Morphological analyzer and generator
- Tamil - Firefox browser, Email
- Telugu - Telugu to Hindi language translation system, Publications,
- Urdu - Word processor cum publisher, encyclopedia, language resources (rare and classic literature), Urdu language design guide and Pascii resources, English version of Open Office

The above software could be used for, word processing**,** transliteration, spellchecking, Optical Character Recognition (OCR) etc.

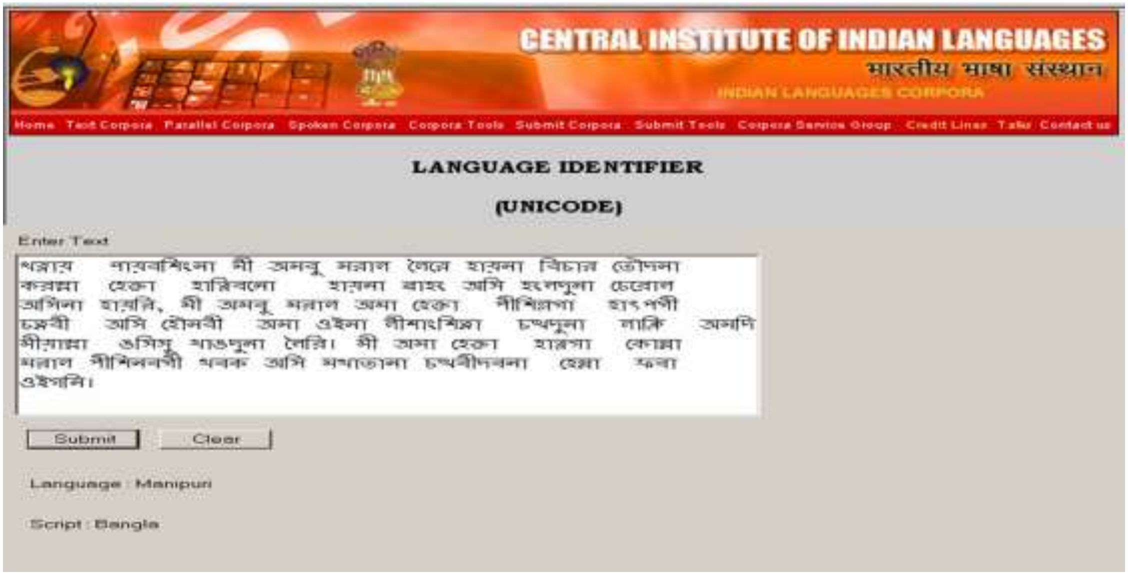
In addition several organizations in India, devoted to NLP and AI research have developed several NLP tools for Indian languages. Some of the NLP tools are:

2.11 Corpora

Plural form of Corpus is the collection of words from writing, be it a novel or a collection of novels or any forms of literature irrespective of any discipline. Corpora databases are created initially in all these languages which form the backbone of any NLP activity. Keeping the uses of corpora all most all the disciplines has developed corpora. The same in Indian languages also does not lag behind. In the present day scenario varieties of corpora are being developed depending on the usability of the same such as Plain corpora, annotated corpora, total vocabulary in a language, Morphology etc., and based on these, several NLP tools are being developed in Indian languages. Some of them which are useful for information retrieval are discussed here.

2.12 Language Identifier

Language identifier plays a very important role in multilingual IRS. In the writing process of Indian languages too, single script is used for many languages (Sharada2006). Unless the language code is entered in the catalogue, the system cannot retrieve the documents related to language search. In a multilingual country like India one cannot expect the indexer to know all the languages. Also if the name of the language in which the book is published is not mentioned any where in the book, one feels the need of automatic language identifier. The following schema exemplifies the language identifier developed by the Central Institute of Indian Languages, Mysore (CIIL)



LANGUAGE IDENTIFIER

(UNICODE)

Enter Text

থগ্রায় শায়বশিংনা মী অমনু মরাল লৈরে হায়না বিচার তৌদনা
করম্না হেক্তা হারিবলো হায়না বাহং অসি হংলদুনা চেরোল
অগিনা হায়রি, মী অমনু মরাল অমা হেক্তা মীশিংনা হাংশগী
চংবী অসি হৌনবী অমা ওইনা মীশাংশিংগা চপ্থদুনা লাকি অমনি
মীয়াম্না ওসিগু শাওদুনা লৈরি। মী অমা হেক্তা হায়না কোম্না
মরাল মীশিংনগী থবক অসি মশাতোনা চপ্থীদবনা হেম্না ফনা
ওইগনি।

Language : Manipuri

Script : Bangla

Schema 1

2.13 Word Frequency Count

The uses of Word frequency count program in IRS are immense. The basic function in preparing the IRS tools is dependent on the Domain Specific Concepts. In addition, lists such as: Stop words, Proper names of Persons and Places, Chronology, etc could be prepared. For experiment sake, the list of document titles in a particular language may be selected and run in the Word frequency count program. A sorted list of words with their frequency is obtained in the target language from which these files could be very easily created.

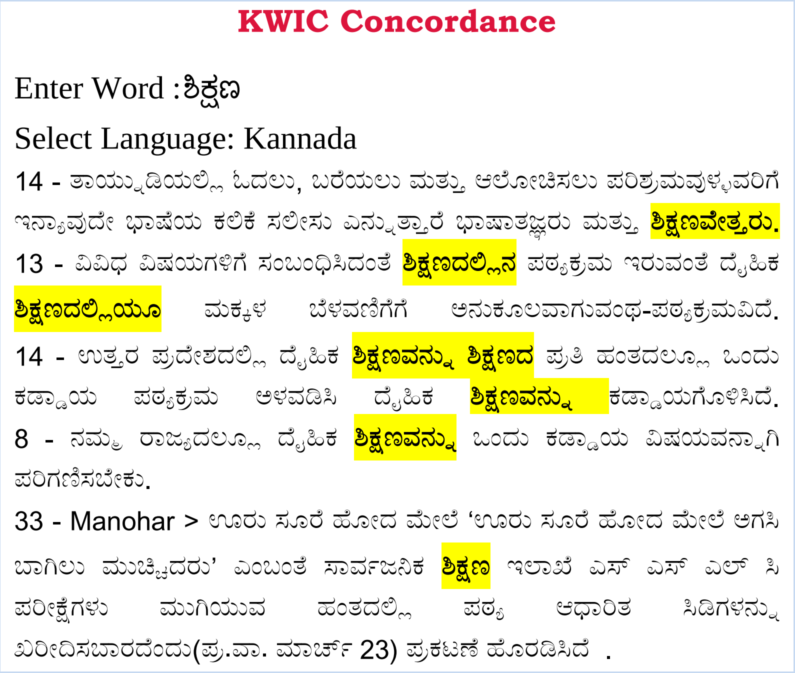
KWIC-KWOC Retriever/Concordance

Key Word in Context (KWIC) and Key Word out of Context (KWOC) are not new phenomena in IRS. In addition the concordance function in NLP tools is a common tool in English but a breakthrough in Indian languages. The word frequency program searches the complete database and gives the output. The selected words are hyper text. The concordance module takes care at the following levels:

- Word with previous word

- Word with next word
- Word with complete title at the sentence level
- Word with all inflections

This module is helpful in information retrieval at both metadata and full text levels. The schema for the last level - Word with all inflections, is given below.



KWIC Concordance

Enter Word :ಶಿಕ್ಷಣ

Select Language: Kannada

14 - ತಾಯ್ನುಡಿಯಲ್ಲಿ ಓದಲು, ಬರೆಯಲು ಮತ್ತು ಆಲೋಚಿಸಲು ಪರಿಶ್ರಮವುಳ್ಳವರಿಗೆ ಇನ್ನಾವುದೇ ಭಾಷೆಯ ಕಲಿಕೆ ಸಲೀಸು ಎನ್ನುತ್ತಾರೆ ಭಾಷಾತಜ್ಞರು ಮತ್ತು **ಶಿಕ್ಷಣವೇತ್ತರು.**

13 - ವಿವಿಧ ವಿಷಯಗಳಿಗೆ ಸಂಬಂಧಿಸಿದಂತೆ **ಶಿಕ್ಷಣದಲ್ಲಿನ** ಪಠ್ಯಕ್ರಮ ಇರುವಂತೆ ದೈಹಿಕ **ಶಿಕ್ಷಣದಲ್ಲಿಯೂ** ಮಕ್ಕಳ ಬೆಳವಣಿಗೆಗೆ ಅನುಕೂಲವಾಗುವಂಥ-ಪಠ್ಯಕ್ರಮವಿದೆ.

14 - ಉತ್ತರ ಪ್ರದೇಶದಲ್ಲಿ ದೈಹಿಕ **ಶಿಕ್ಷಣವನ್ನು ಶಿಕ್ಷಣದ** ಪ್ರತಿ ಹಂತದಲ್ಲೂ ಒಂದು ಕಡ್ಡಾಯ ಪಠ್ಯಕ್ರಮ ಅಳವಡಿಸಿ ದೈಹಿಕ **ಶಿಕ್ಷಣವನ್ನು** ಕಡ್ಡಾಯಗೊಳಿಸಿದೆ.

8 - ನಮ್ಮ ರಾಜ್ಯದಲ್ಲೂ ದೈಹಿಕ **ಶಿಕ್ಷಣವನ್ನು** ಒಂದು ಕಡ್ಡಾಯ ವಿಷಯವನ್ನಾಗಿ ಪರಿಗಣಿಸಬೇಕು.

33 - Manohar > ಊರು ಸೂರೆ ಹೋದ ಮೇಲೆ ‘ಊರು ಸೂರೆ ಹೋದ ಮೇಲೆ ಅಗಸಿ ಬಾಗಿಲು ಮುಚ್ಚಿದರು’ ಎಂಬಂತೆ ಸಾರ್ವಜನಿಕ **ಶಿಕ್ಷಣ** ಇಲಾಖೆ ಎಸ್ ಎಸ್ ಎಲ್ ಸಿ ಪರೀಕ್ಷೆಗಳು ಮುಗಿಯುವ ಹಂತದಲ್ಲಿ ಪಠ್ಯ ಆಧಾರಿತ ಸಿಡಿಗಳನ್ನು ಖರೀದಿಸಬಾರದೆಂದು(ಪ್ರ.ವಾ. ಮಾರ್ಚ್ 23) ಪ್ರಕಟಣೆ ಹೊರಡಿಸಿದೆ .

Schema 3

2.14 Transliterator

Usual practice in developing bibliographic database for books in Indian languages is preparation of bilingual metadata comprising of the data in original language and its English transliteration. For the purpose of transliteration, varieties of software are available in the IT market. Freely available online dictionaries including famous search engine Google provide transliteration in major Indian languages.

- **Automatic Transliterator for Indian Scripts**

The automatic transliterator developed by Linguistic Data Consortium for Indian Languages (LDC-IL), CIIL has the standard transliteration scheme for the Indian Languages. And this being a linier scheme as opposed to sub-script/super-script with automatic transliteration software, comes handy for digital information data entry and retrieval. Multilingual, multi-script contexts demand inter-convertibility between scripts. It renders any Indian language Unicode data into standardized English transliteration scheme specifically evolved for the purpose and at the same time it does backward transliteration from English to any Indian script. Not only this, but also source language from Devanagari, Kannada, Malayalam, Tamil and Telugu to any of these Indian languages. Some of these works are in progress.

2.15 Lexical Interface with the Available Web Resources

Lexical interface acts as a complementary tool in an IRS. If such a tool is made available in the library online public access catalog (OPAC), any doubt related to meaning of a concept could be got cleared immediately. The best example is WordNet. Lexical interface for Indian language documents shall be Indo-WordNet which is now available online for Hindi and Marathi languages. For other Indian languages it is in the developing stage.

2.16 OCR for Indian languages

Digitization is the buzz word in modern libraries. There is a field in the library OPAC to house the digitized full text along with the meta data which make the IR process still user friendly. Suitable tool for digitizing the documents in Indian languages is a challenging issue. While digitizing books in Indian languages, Optical Character Recognition (OCR) is very much essential. But due to the characteristics of Indian scripts, the OCR systems that work for English is not applicable for Indian languages.

Further, in Indian context, most of the works have been carried upon for the OCR for Devanagari, Bangla (Chaudhuri, B.B. and Pal, U, 1997 & 1998) and Telugu scripts (Atul Negi, 2001) and for the rest it is still in a developing stage, for example, Kannada (Kunte, Sanjeev). Based on the research done by IIIT, Hyderabad, OCR work on Oriya and Bengali hand-written character recognition is in progress (Vamshi Ambati).

2.17 Information Extraction

Information extraction is one of the NLP features. A grammar entitled 'Approximate Grammar' is proposed by IIIT, Hyderabad and explains how it can be used to extract information from a document. It states that, as the structure of informational strings cannot be defined well in a document, we cannot use the conventional grammar rules to represent the information. Hence, the need arises to design an approximate grammar that can be used effectively to accomplish the task of Information extraction. (**http://arxiv.org/ftp/cs/papers/0305/0305004.pdf**

2.18 Online Thesaurus

As per CDAC (http://iplugin.cdac.in/nlp.htm) Thesauri which provide much more semantic information than dictionaries are a vital tool for search-engines, data-mining and information retrieval. GIST has created a Thesaurus Building Engine, which will ensure that the structure of the thesaurus with its **hyponyms** is correctly indexed. This will result in a fast thesaurus and quick information retrieval.

2.19 Information retrieval thesaurus(IRT)

The information retrieval thesaurus helps in formulating search expressions enabling, in relation to the user's query. Such an IRT in all the Indian languages will be an ideal addition to the multilingual digital libraries.

An IRT is used for vocabulary management:

- to enable the use of standard terms in indexing, formulating search expression and retrieval,
- to find an appropriate standard term (descriptor) for a given concept, and
- to find and select terms related to the descriptor.

In addition, many NLP techniques, including stemming, part of-speech tagging, compound recognition, de-compounding, chunking, word sense disambiguation and others, have been used in Information Retrieval to certain extent.

3 Analysis

From the sample data analysis is done based on the title of the document. In some cases full text was available and made the work easy. But in some cases bibliographical data was available. Of course in the journal databases abstract was available. The articles are classified according to topic wise. The complete year wise list has been appended to this paper and name of the authors have been copied as they are from the sources giving importance to each of the contributors. Since this study will be made available online each part of the name would be a search field.

3.1. NLP Application to IRS

Following are the articles that deal with NLP and IRS. They have adopted different methodologies. Some papers though general could be adapted to Indian languages also with slight modification depending on the language structure. It is observed that both syntax and semantics have been taken into consideration. The titles are in italics.

i. '*A System for Personalized Information Retrieval based on Domain Knowledge'* was presented in 2003 irrespective of any language but could be applied to any language.

ii. *'Bengali and Hindi to English CLIR (Cross-Lingual Information Retrieval) Evaluation'* was discussed in 2008.

iii. '*A Personalized Information Retrieval Module for Retrieving Learning Materials' was presented in 2009.*The three studies a – c were done by the same author with different associates.

iv. *Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method* presented in 2007 is a very important study as far as Indian languages are concerned. The IRS method discussed here could be applied to other Indian languages also.

v. *Semantic information retrieval from enterprise data* was discussed in 2010 irrespective of any language.

vi. In the present multi modal access to information, the paper *Information retrieval using text detection from video by image processing and speech recognition for video search optimisation* was discussed in 2010.

vii. In order to narrow down the search and provide pin pointed answer to the query semantics is playing a vital role. Such a presentation in 2010 was *Knowledge management with semantic annotation: An approach for information retrieval using KIM.*

viii. *A multi stage fall-back search strategy for cross-lingual information retrieval* was presented in 2005.

ix. *NLP and its Application in Information Retrieval with Special Reference to Kannada* was a practical application paper in 2005.

x. In 2006, *Multilingual information retrieval system and Unicode -Digitizing multilingual libraries: a case study at CIIL d*iscussed NLP tools such as language identifier, Transliterator, concordance, lexical interface, etc., with examples and their uses in IRS.

xi. In 2007, the study *Information retrieval in Indian languages: a case study of plural resolution in Telugu language* highlighted the IRS in Telugu.

xii. *An Information retrieval Approach Based on Semantically Adapted Vector Space Model* in 2009 discussed IR based on a specific NLP model.

xiii. The paper in 2006 *Multilingual Information Access: Information Retrieval and Translation in a Digital Library* discuss the present day need in IR related to Multilingual Information and translation.

xiv. The study in 2009 *STAIR:A System for topical and aggregated information retrieval* is the out come of Intelligent Human Computer Interaction

xv. *Natural Language Processing and Information Retrieval* is a book published in 2008 which discusses many issues related to the topic

xvi. *Three Stage Refinement Method of Information Retrieval with application to newspaper articles and online documents* presented in 2001 could be applied to newspapers in Indian languages since most of them are available in the internet

xvii. *Digital Libraries in Multilingual Countries: An Indian Case study* in 2008 discuss this topic in detail with examples in Indian languages. This paper also discuss topics such as

- “automatic information retrieval of documents in Indian languages in pictorial as well as hypertext form”(**http://www.isical.ac.in/~cvpr/index.html**)
- .Multilingual information retrieval in Indian languages (Vamshi Ambati).

xviii. *Bhasa: A Corpus-Based Information Retrieval and Summariser for Bengali Text*. Bhasa, is a corpus-based search engine and summarizer that performs document indexing and retrieves information based on key words using vector space retrieval method. Authors states that Basha-Search may not be as successful as we expected but we are looking forward to do more research in this area and extending our first retrieval engine to solve real life problems and bring Basha-Search as one of the successful retrieval engine.

xix. *WebKhoj: Indian language IR from multiple character encodings*, presented in 2006 states that more than 95% of Indian language content on the web is not searchable due to multiple encodings of web pages.

xx. *Hashing-based approaches to spelling correction of personal names*, presented in 2010 propose two hashing-based solutions to the problem of fast and effective personal names spelling correction in People Search applications. The key idea behind this method is to learn hash functions that map similar names to similar (and compact) binary codes.

xxi**.** *Multilingual people search* claims that, People Search is an important search service with multiple applications with the proportion of non-English users on a steady rise, people search services are being used by users from diverse language demographics. Users may issue name search queries against these directories in languages other than the language of the directory, in which case the present monolingual name search approaches will not work. It gives a demo that present, a Multilingual People Search system capable of performing fast name lookups on large user directories, independent of the directory language.

xxii. The papers *"A term is known by the company it keeps": On Selecting a Good Expansion Set in Pseudo-Relevance Feedback* and *"They Are Out There, If You Know Where to Look": Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval* discusses pseudo-relevance feedback (PRF) and Cross-Language Information Retrieval (CLIR) respectively that improves the retrieval performance of Information Retrieval (IR) systems in general.

xxiii. *Experiments in CLIR using fuzzy string search based on surface similarity* in 2009 discusses Cross Language Information Retrieval (CLIR) between languages of the same origin. And the paper *An Unsupervised Approach to Product Attribute Extraction* the authors claim that they are able to achieve 92% precision and 62% recall in their experiments. The paper in 2010 *Evaluation of English-Telugu and English-Tamil Cross Language Information Retrieval System using Dictionary Based Query Translation Method* states that the CLIR helps the users to pose the query in one language and retrieve the documents in another language. In this system, the query in English shall retrieve the documents in Tamil and Telugu.

xxiv. *Tamil text analyser* in 2003 claims that chunking can be used to retrieve the information from the documents depending on the chunks rather than words. Nouns and Noun phrases are more useful for retrieval and extraction purpose.

3.2. Language Identifier

In 2002 *Script identification in printed bilingual documents* was published for Tamil
In 2004 discussed '*Language Identification from Small Text Samples*' for the languages Hindi, Bengali, Marathi, Punjabi, Oriya, and Telugu.

In 2006 '*Language identification from small text samples'* has been discussed in general.

The same in 2010 '*Addressing Challenges in automatic language identification of Romanised text'* for the languages Hindi, Telugu, Tamil, Kannada & Malayalam was presented.

The four major languages under the Dravidian language family have language identifier software. It is found from one of the studies that even dialect like Konkani could be identified.

3.3 Transliterator

a. *A More Discerning and Adaptable Multilingual Transliteration Mechanism*

b. Multilingual Information Access: Information Retrieval and Translation in a Digital Library

c. *Multilingual information retrieval system and Unicode*

d. *Digital Libraries in Multilingual Countries: An Indian Case study* in 2008 discusses the automatic transliterator in Indian languages with examples in Bengali, Hindi, Dogri and Maithili.

3.4. OCR for Indian languages

General:

The paper *Digital Libraries in Multilingual Countries: An Indian Case study* in 2008 discusses in general the aspect of OCR in Indian languages with examples.

An OCR architecture for indexing Indian language based web documents was discussed in a seminar which was not pertained to any single language in 2005.

Kannada: In 2000 a study *A font and size independent OCR for printed Kannada using SVM* was taken in fulfillment of ME Project Report, Indian Institute of Science, Bangalore. This required further improvements.

Then, **i**n 2002, *A font and size-independent OCR system for printed Kannada documents using support vector machines* was published.

A simple and efficient optical character recognition system for basic symbols in printed Kannada text was published in 2007.

Gurmukhi: *A Gurmukhi script recognition system* was discussed in a seminar in 2000. *A post-processor for Gurmukhi OCR* was published in 2002.

Hindi: The topic *A complete OCR for printed Hindi text in Devanagari script* was discussed in a seminar in 2001.

In 2002 *Neural network based system for script identification in Indian documents* worked on Hindi and Kannada.

Tamil:

In 2000 *Optical character recognition for printed Tamil script* was the Master's thesis, Indian Institute of Science, Bangalore

A complete OCR for Tamil printed text was discussed in 2000 in a seminar at Singapore.

Telugu: *An OCR system for Telugu* was discussed in a seminar in 2001.

Oriya: *Automatic recognition of printed Oriya script* got published in 2002

In 2006 *Handwriting segmentation of unconstrained Oriya text* was published.

Bengali: *A high performance domain specific OCR for Bangla script*. The authors state that "the entire technique significantly increases the performance of the OCR for a specific domain to a great extent".

3.5 Spell Checker

Design and implementation of a Spell Checker for Assamese and *Spell Checking in MSWord for Assamese*

Digital Libraries in Multilingual Countries: An Indian Case study in 2008 discuss this aspect with examples in Kannada, Bengali, Oriya and Marathi

3.6 Lexical Interface

Exploring Hindi WordNet as a lexical interface and subject headings tool in library OPAC presented in 2010 discusses the uses of Indo-wordnet as an IRS tool with special reference to lexical interface in the library OPAC. Few related papers are:

Lexical chains as document features, 2008

*Using Semantics in document representation: A Lexical chain approach,*2009

*Document clustering using lexical chains.*2007

*Text Categorization in Indian Languages using Machine Learning Approaches.*2007

3.7 Information Extraction

Following are some articles in this topic:

Template based information extraction and subsequent template matching, was discussed in 2001

Knowledge extraction from Indo-Aryan family of natural languages using a rule based approach was discussed in 2002.

Rapid domain porting of an intermediate level information extraction engine was discussed in 2003.

Heuristic Acronym Extraction Using Linguistic Features and *Corpus Level Information Extraction* were discussed in 2004.

Dissertations Abstract International published *Knowledge-based methods for automatic extraction of domain-specific ontology* in 2007, which could be customized for other languages.

Automatic key phrase extraction from Bengali documents was discussed in 2010.

3.8 Information retrieval thesaurus

1. *On Automatic Construction of a Thesaurus* was discussed in 2004.

2. *Bilingual Tamil-English Bilingual Thesaurus for Use in Document Indexing and Retrieval* was published in 2007. This is a thesaurus to assist document indexing and information retrieval. Information Retrieval Thesaurus (IRT) differs from a conventional language thesaurus, such as, Roget's *Thesaurus of the English Language*. It helps in selecting an appropriate term for a given concept for which the user may have only a vague idea about a suitable term.

4. Indian Language wise contribution

From the sample data if individual languages are examined following Table depicts the result.

Table 2: Language wise contribution

Rank	Language	Contribution
1	Bengali	38
2	Hindi	32
3	Tamil	17
4	Telugu	15
5	Kannada	14
6	Oriya	8
7	Assamese	6
8	Malayalam	6
9	Gujarati	3
10	Gurumukhi Script	3
11	Marathi	3
12	Punjabi	3

Through this paper many useful studies and NLP achievements in Indian languages aiding in information retrieval system have been brought to light. Some general studies too could be used for IRS in other Indian languages with slight modification depending on the structure of the target language. At the same time the languages where seldom research is done in this area are also shown.

Some of the NLP application studies are amazing. Few such studies look incomplete and need improvement because they are done for the demonstration of the basic technology based on minute data and cannot be used for application at the concrete level. Hope in the near future many of these will be developed into usable application since many of the organizations are getting support and coming out with useful NLP tools in major Indian languages.

One more suggestion is that all such institutions come forward together and start developing or improving the existing NLP tools duplication could be avoided. Very important is that whatever they develop it should be brought to the notice of the public so that they can get the proper feedback for improvement. This is the best way we could see the development of a language as well as its application in information retrieval in a proper perspective.

Note: In the annexure to this, papers that are discussed here are listed. Nearly ten papers are incomplete. Authors may help to complete this database. Also users are requested to send the details of their contributions if they are not included here to sharadamallik@gmail.com.

==

References

1. Chaudhuri, B.B. and Pal, U. “An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari (Hindi).” In Proceedings of 6th International Conference on Document analysis and Recognition, 1997.
2. Chaudhuri, B.B. and Pal, U. “A complete printed Bangla OCR system”. Pattern Recognition, 31, 1998.
3. Google Indic Transliteration **http://www.google.com/transliterate/indic**
4. Sharada, B. A. 2006.Multilingual information retrieval system and Unicode - Digitizing multilingual libraries: a case study at CIIL. ICDL 2006. New Delhi: TERI. 422-436.
5. Sharada, B. A. 2010.Exploring Hindi wordNet as a Lexical Interface and Subject headings Tool In Library OPAC.2010 In Principles, Construction and Application of Multilingual Wordnets: Proceeing of the 5th global Wordnet Conference New Delhi: Narosa Publishing House.62-69.
6. Sharada, B. A. 2010. *Indian linguistia and Information Science*. Mysore: Central Institute of Indian Languages.
7. Thorsten Brants. Natural Language Processing in Information Retrieval. Google Inc. citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87…
8. ***http://www.isical.ac.in/~cvpr/index.html***

===

B. A. Sharada, Ph.D.
Retired Librarian
Central Institute of Indian Languages
Ministry of Human Resource Development
Manasagangothri
Mysore 570 006
Karnataka
India
sharadamallik@gmail.com