Natural Language Processing Activity

Center for Indian Language Technology,
Computer Science and Engineering Department,

at

Indian Institute of Technology Bombay

Introduction:

Center for Indian Language Technology (CFILT) was set up with a generous grant from the Department of Information Technology (**DIT**), Ministry of Communication and Information Technology, Government of India in 2000 at the Department of Computer Science and Engineering, IIT Bombay. Prior to this the Natural Language Processing (NLP) activity of the CSE Department, IIT Bombay took off in 1996 with a grant from the United Nations University, Tokyo to create a multilingual information exchange system for the web. The project called **Universal Networking Languag**e (UNL; www.undl.org) was participated in by 15 research groups across continents.

At any point of time about 30 research members work in CFILT, which includes PhD , masters and bachelor students, faculty members, linguists and lexicographers.

Research Themes:

Deep semantics and multilinguality has throughout played a pivotal role in the activities of CFILT. The stress on semantics has led to research in the following fronts:

- **Lexical Resources:** Multilingual Wordnets and ontologies and their linking
- **Lexical and Structural Disambiguation:** Resolve word and attachment ambiguities
- **Shallow Parsing:** Identifying correct parts of speech, named entities and non-recursive noun phrases for Marathi and Hindi
- **Cross Lingual Information Retrieval:** Indian language query to English and Hindi Retrieval
- **Machine Translation:** Automatic translation involving Marathi, Hindi and English
- **Text Entailment:** Testing if a piece text (hypothesis) is inferable from another (text)
- **Sentiment Analysis:** Detecting polarity- positive/negative/neutral- of a given document, especially reviews

Leadership in National Endeavors:

The NLP group at IIT Bombay has for long been involved in large scale NLP initiatives funded by the Ministry of Communication and Information Technology, India and also Ministry of Human Resource Development. Leading academic institute and industries of India are involved in these efforts, *viz.*, *IIT Kharagpur, IIIT Hyderabad, CDAC Noida, CDAC Pune, Jadavpur University, IIIT Allahabad, Indian Statistical Institute, Guwahati University, Manipur University, Assam University, Dravidian University, Goa University, Amrita University, University of Hyderabad, Thapar Institute* and *Punjabi University*.

- **Consortium Project on Cross Lingual Information Access (CLIA):** Service user queries in *Bengali, Hindi, Marathi, Punjabi, Tamil* and *Telugu* by retrieving documents in Hindi and English and displaying content in the query language.
- **Consortium Project on Indian Language to Indian Language Machine Translation (ILILMT)**: Machine Translation between *Hindi* on one hand and *Bengali, Marathi, Punjabi, Tamil, Telugu* and *Urdu* on the other hand.
- **Consortium Project on English to Indian Language Machine Translation (EILMT):** Machine Translation between *English* on one hand and *Bengali, Hindi, Marathi, Oriya, Tamil* and *Urdu* on the other hand.
- **Consortium Project on Building IndoWordnet**: Hindi Wordnet is already released for public download. Marathi Wordnet is reaching maturity. The Wordnet building efforts include **North-East Wordnet** involving Assamese, Bodo, Manipuri and Nepali, **Dravidnet** involving Kannad, Malayalam, Tamil and Telugu and **Indradhanush** involving Bangla, Gujarathi, Konkani, Punjabi and Urdu. All these words are being linked to Hindi and English leading eventually to IndoWordnet. The Linguistic Data Consortium for Indian Languages (*LDC-IL*) at Central Institute of Indian Languages (*CIIL*) has funded creation of **Sanskrit Wordnet**.

Collaboration with Industry:

Tata Consultancy Services (**TCS**), the leading software industry of India, has long been our collaborator in research on deep semantics (UNL). America Online (**AOL**) has set up joint research program on Sentiment Analysis. HP Laboratories (**HP Labs**) have funded activities on ontologies. **Xerox** India and CFILT will work on parsing technologies. Research Grant was obtained from **Microsoft Research India** for Multilingual database creation based on Hindi Wordnet. **IBM India** initiated research collaboration for Unstructured Information Management.

Role in Providing Standards:

A multilingual dictionary standard provided by IITB has been adopted in India to create dictionaries of different languages starting with Manipuri in the East to Marathi in the West and Kashmiri in the North to Tamil in the South.

Senses	Hindi	Marathi	Bengali	…	Tamil
(sun)	(सूय, सूरज, भानु, भा क, भाकर, दनकर)	(सूय, भानु, दवाक, भा क, र, व, दनेश, दनमणी)	…	…	…
(cub, lad, laddie, sonny, sonny boy)	(लड़का, बालक, ब च, छोकड़ा, छोरा, छोकरा)	(मुलगा, पोरगा, पोर, पोरगे)	…	…	…
…	…	…	…	…	…

Figure 1: Multilingual Dictionary Standard adopted in India

The unique feature of the standard is the linkage amongst concepts (**synsets**) and not amongst individual words.

Visibility of CFILT:

The Hindi Wordnet project team was awarded the *P. K. Patwardhan Award for Technology Development* of IIT Bombay in 2009. The initiative also won the *Manthan Award 2009* (further information on: http://manthanaward.org/section_full_story.asp?id=865) for ICT for development.

Faculty members have been recognized with important offices in top NLP fora like ACL and COLING (Prof. Pushpak Bhattacharyya), President's Medal for contribution to Sanskrit (Prof. Malhar Kulkarni) and so on. Periodic news and TV coverage of activities is CFILT is a regular feature.

Impact and Uses of Hindi Wordnet:

- Free download with API under *GPL*
- Available from LDC (linguistics data consortium), Upenn
- To be available from ELRA: language data repository of Europe
- Available from LDC-IL: LDC of India
- Daily reference from all over the world
- More than 281000 hits so far since 2006
- More than 5000 downloads
- Pivot for Wordnets of many Indian languages

- Base resource used by many researchers for IL work on translation, summarization, cross lingual search
- Commercial license acquired by major search engines companies

Important conferences, symposia and workshops have been organized by CFILT in the past:

- Summer workshop on Ontology, NLP, IE and IR, IIT Bombay, India, July, 2008
- Symposium on Modelling and Shallow Parsing of Indian Languages, IIT Bombay, India, April, 2006
- Universal Networking Language Summer School, Mumbai, India, April, 2003
- International Conference on Universal Knowledge and Languages, Goa, India, November 2002

Contact:

Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering,
Indian Institute of Technology Bombay

Powai, Mumbai 400 076. INDIA
Telephone No.: (+91-22) 25764729, 7718(Office)
Fax No.: (+91-22) 25723480, 0290
E-mail: pb@cse.iitb.ac.in
Website: http://www.cfilt.iitb.ac.in

Relevant URLs

www.cfilt.iitb.ac.in for resources
www.cse.iitb.ac.in/~pb for publications