Since UNL has generated lot of interest in the research community for natural Lang a means for exchange of information on the internet in natural languages. He acknowledged the key role played by three very special individuals in pioneering the UNL project- Dr. Hiroshi Uchida, Dr. Kazuhiko Nishi and Dr Della Senta. He explained the problems faced by the United Nations (UN) related to the "Language Barrier". There are 186 major languages spoken throughout the world out of which six are the official UN languages. The division of UN which is closest with respect to its functions to Information Technology is International Telecom Union (ITU). ITU spends 40% of its annual budget to ensure that its products are translated in all six official UN languages. To overcome this language barrier, Universal Networking Language (UNL) was conceived.

Gradually Hindi also was brought into the umbrella of this project, as India is a country with vast population and great language diversity. Prof Menon further added that, at present, there exist mainly three types of divides in this world. The first divide is the North-South divide, i.e., Rich and the poor divide. The second is the Digital Divide, i.e., the divide between people who have access to the Internet and those who do not. The third divide is the language divide due to the dominance of English language on the Internet. The challenge faced by UN today is how to ensure that these divides break down. Help is needed from commercial centers to take it down the line to the masses. The idea is to have it in public domain and not to make it a commercial property.

Dr. Uchida explained in detail the concept of the Universal Networking Language (UNL). The three pillars of this language are the Universal Words (standing for concepts), the Relation Labels (standing for relationships among the concepts as they appear in a specific sentence) and attribute labels (standing for speaker's attitude with respect to tense, number, mood etc.). The processes of "Deconversion", i.e., generation of natural language from UNL and "EnConversion", i.e., analysis of natural languages into UNL were explained . These two put together lead to (semi) automatic machine translation.

Prof. A. Sen and Prof. S. Sanyal described the ongoing work on speech and language technology in TIFR. They talked about the tools developed for speech processing at TIFR related to segmentation, speaker dependent phoneme identification, speech database etc. It was clear that today's speech research has to draw a lot from the natural language processing research.

Mrs. Tessie George, narrated what a professional translator expects from a machine translation software.

Dr. Mudur gave an overview of various activities like encoding techniques and standards, multilingual text processing and font design, OS level support for localization, Intelligent Information Retrieval and Machine Translation. Mr. Durgesh Rao, NCST, talked at length about Information Retrieval, Machine Translation and Language Engineering related work at NCST, particularly the English-Hindi translation tool, MaTra.

Dr. Sengupta described the ongoing research activities on language and speech technology at Tata Infotech Ltd. i.e creation of speech database was one of TIL's focus.

The last session include the presentations and demonstrations by the UNL Research group of IIT Bombay on construction of Hindi-UW dictionary, EnConversion (analysis) of English text to UNL , Hindi EnConversion process, Hindi DeConversion (generation), Marathi DeConversion, building the Indo Word net, Ontology building, UNL based inferencing, Automatic Hyper Linking using UNL, UNL-ML, i.e., XMLification of UNL. Each of these presentations was followed by small demonstrations.

Indo-UK workshop on Language Engineering for South Asian Languages (LESAL), on April 23-25, 2001 at NCST, Mumbai

An Indo-UK workshop on Language Engineering for South Asian Languages (LESAL), was organized at the National Centre for Software Technology, Mumbai, on April 23, 24 and 25, 2001. The aim of the workshop was to stimulate interaction among the Natural Language Processing communities in the UK and South Asia, for the mutual benefit of both communities, and for the state of the art in Language Engineering in the languages of South Asia.

The workshop covered a wide spectrum of NL activity, including corpus linguistics, lexicography, encoding standards, language engineering tools, and application areas such as machine translation, speech processing, OCR and spell checking.

The major suggestions / recommendations were as follows.

- Linguistic Data: There is scope for consolidation of current efforts, and joint work in building parallel corpora for English and Indian languages. Some of the possible collaborators and sources of parallel corpora include CIIL, CIEFL, the National Book Trust, the Sahitya Academy, news agencies, media and government bodies. This could aid further work in lexical resource building, for example, the LERIL initiative for development of Lexical Resources in Indian Languages. Speech corpora, and corpora for Indian English are other lacuna that need to be filled. Tribal languages may also benefit from such efforts to the extent possible.

- Tools and Environments: There are very few shareable tools and environments for Language Engineering. One example is the GATE (General Architecture for Text Engineering) effort at Sheffield. There is scope to work together in adding South Asian language support into such tools and environments, and also to broaden the scope of such tools, by adding components specific to these languages, and also application areas such as translation and generation. Evaluation and comparative study of existing tools and environment should be made.

- Applications: The major application areas of Language Technology are Information Retrieval and Extraction, Translation aids, OCRs, spellcheckers and speech processing systems. Some of the lesser known, but very relevant applications include localization, language learning, and communication tools for the blind, the illiterate and the semi-literate. The specific modes of cooperation in these areas needs to be explored further.

- Standardisation: There is an urgent need for India to actively promote and participate in standards building activities. There are four aspects or levels of standardization — Encoding (characters, fonts etc), Representation of text (logical organization), Linguistic Annotation, and Access (indexing and metadata). India should participate in global initiatives such as the TEI (Text Encoding Initiative) and OLAC (Open Language Archive Community), and make sure that its unique language standards needs are addressed, since it is not necessary that methods used for other languages are directly applicable to South Asian languages. For example, there are some unique features of South Asian languages, such as their phonetic nature and presence of composite glyphs, which makes it necessary to handle them specially at the encoding level. Similarly, due to their rich morphology and relatively free word order, they may need different kinds of analysis, representation and tagging models.

- Evaluation: Systematic and objective evaluation of NL technologies and products, though not easy, is a necessary and effective mechanism to establish and extend the state of the art. Evaluation can take three forms, depending on how mature and established a technology is

  a. Technology Oriented Evaluation, which typically involves voluntary participation of groups with focus on specific components or techniques, not necessarily complete products.

  b. Product Oriented Evaluation, which involves user feedback on actual systems.

  c. Certification by some authority as to interoperability, usability etc. It was felt that Indian NL groups should get together and try to organize Technology Evaluations, along the lines of the TREC and MUC conferences.

New Software Testing Facility

Under Soft QC/QM project of Computer Development Division of Ministry of Information Technology Electronics Test & Development Centre, Thiruvanmiyur (Near Lattice Bridge), chennai-600041 has established a software Testing & Training Laboratory (STTL) at chennai. This laboratory is capable of testing software products as per ISO/IEC 12119 standard. Academia, Industry and Language Technology projects being carried out under MIT shall benefit from this facility.

Contact Person :

Mr. R. Subburaj, Director ETDC, Chennai

E-mail : etdccn@md3.vsnl.net.in Tel : 044-4928192

Now Domain Names in Regional Languages

- US based domain name Registration Company, Network Solutions (http://www.networksolutions.com) announced domain name registration service in nine Indian regional languages. They are launching domain name registration services in nine different Indian Languages-Bengali, Gujrati, Hindi, Kannada, Malyalam, Oriya, Punjabi, Tamil and Telugu to enable people to break the language barrier. Besides bridging the digital divide the expansion of multilingual services would induce small and medium businessmen to conduct transactions online in their language. Until recently every domain name registration was done using exclusively ASCII characters. India ranks third in the domain name registration in the Asia-Pacific market after Korea and China. Along with these languages one can use .org, .net and .com for Internet Domain Name Registration. The invitation price for the service initially would be $25.

- VeriSign Global Registry services (http://www.verisign.com) which maintains the master database for the majority of the World's domain names has decided to accept names in the simplified and traditional Chinese, Japanese, and Korean characters. Next, VeriSign plans to accept domain names in Arabic, and names with Spanish and Portugese characters not used in the English.