Natural Language Processing Activities in CDAC, Noida

V. N. Shukla

Abstract - Natural language processing is a field of computer science and Computational linguistics and is associated with human – Machine interaction having two major components 1) Natural language generation systems to convert information from computer to natural language and 2) Natural language understanding systems to convert reverse way. As such NLP is the basic interface between Human and Computer. Many NLP problems fall under both generation and understanding

NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of artificial intelligence. The term natural language is used to distinguish human languages (such as English, Hindi etc.) From formal or computer languages (such as C++, Java) NLP may encompass research and development in both text and speech.

While International scenario shows a struggle in NLP research at many Institute of rapport, Indian scenario is more complex than all of them put together, because not only there are 22 official languages with as many dialects as 1028, but there are only 11 scripts to represent these languages. Complexity increases further because of similarity and dissimilarity between these Indian languages. It's true that all Indian Languages (ILs) have emerged out of core ancient language Sanskrit and mostly all of them follow Paninian grammar, but that itself is a problem as their dependence on Sanskrit and Panini is varying.

CDAC being a premier research National Institute has mandate of carrying out research and develop utilities as products for common masses in Multilingual Computing. Almost all centers of CDAC are having NLP labs where research work is continuing since their inception. Some prominent and visible activities are going on at Pune, Mumbai, Noida, Thiruanathpuram, Kolkata and Banglore. These activities involves development of smaller utilities like Desktop and internet access in Indian languages, and core research in areas of Machine Translation, Optical Character recognition, Cross Lingual access, Search Engines, standardization, Digital Library etc. The relationship with International Institutes and standardization bodies has been established to uplift activities to international standards and avoid reinventing the wheel. This paper describes NLP activities in CDAC Noida and their relevance with the need at National level.

Keywords: Multilingual Computing, Natural Language Processing, Indian Languages, OCR, Machine Translation, CLIR, IR

1. Introduction

NLP is the basic interface between Human and Computer. It is although considered as a branch of Artificial Intelligence; it is the need for almost all IT applications at the front end at least. It is also considered as major localization problem. No software application can proliferate to all users unless it has utility to operate it with local language. Particularly in India where only 3 % population is considered to know English. While literacy is increasing, computer literacy can increase only when language interface is available to

operate that package. Even Machine translations, OCRs, CLIRs are needed to increase the penetration of IT, since the best of the system can be used only when not only the interface but functionality also matches with the thought process of user, which often is governed by the language user speak and understand.

Many NLP problems fall under both generation and understanding e.g. a computer needs morphology model to understand an English sentence, and a model of morphology is also needed for producing a grammatically correct English sentence. NLP overlaps significantly with computational linguistics. Natural language term is introduced to distinguish human languages (such as English, Hindi etc.) from formal (computer languages such as C++, Java etc). Usually R & D in NLP in both text and speech involves some typical tasks. Major among these are:

- Automatic summarization
- Language reading and writing aids
- Information extraction
- Information retrieval (IR)
- Machine translation
- Named entity recognition (NER)
- Natural language generation
- Natural language search
- Natural language understanding
- Optical character recognition
- Anaphora resolution
- Query expansion
- Question answering
- Speech recognition
- Spoken dialogue system
- Stemming
- Text-to-speech

2. Evaluation of natural-language-understanding systems

AI goal initially was to give computer the ability to parse natural language sentences similar to sentence diagrams that grade-school children learn. One of the first such systems was developed in 1963 by Susumu Kuno of Harvard. The system revealed the depth of ambiguity in English language.

Natural languages have very little inflectional morphology to distinguish between parts of speech, additional information is implied by emphasizing them by speech. E.g. "I never said she stole my money" can be spoken in different ways causing inherent difficulty for NL processor in parsing it. (Someone else said it, but I didn't / I said someone took it; I didn't say it was she / I said she stole something of mine, but not my money / etc.). Because of such inherent problems in NLP, the definition of proper evaluation criteria is often considered the best way to specify an NLP problem. The goal of NLP evaluation is to measure one or more qualities of an algorithm or a system, in order to determine whether and to what extent the system answers the goals of its designers, or meets the needs of its users. Research in NLP evaluation has received considerable attention.

The first evaluation project on written texts was in 1987 (Pallet 1998). Thereafter the Parseval/ GEIG project compared phrase-structure grammars (Black 1991). Tipster project

evaluated on tasks like summarization, translation and searching (Hirshman 1998). Morpholympics (1994) compared German taggers. Senseval and Romanseval campaigns checked for semantic disambiguation. Sparkle campaign (1996) compared syntactic parsers in English, French, German and Italian. Grace project (France, 1997) compared a set of 21 taggers for French (Adda 1999). 13 parsers for French were compared in Technolangue /Easy project in 2004. Large-scale evaluation of dependency parsers was also performed in 2006 and 2007. In Italy, the Evalita campaign was conducted in 2007. In ANR-Passage project (France, 2007) 10 parsers for French were compared. CDAC Noida has also prepared a test bed. It is a semi-automatic testing work bench for testing English to Hindi Machine translation system and used for evaluating various approaches being tried in India under the ages of DIT in 2004. Depending on the evaluation procedures, NLP evaluation could be categorized as: Black-box vs. Glass-box evaluation and Automatic vs. Manual evaluation.

In Black-box evaluation the NLP system is run on a given data set and various parameters related to the quality of the process (speed, reliability, resource consumption) are measured. Glass-box evaluation looks at the design of the system, the algorithms that are implemented, the linguistic resources it uses etc. Due to the complexity of NLP problems, it is difficult to predict performance only on the basis of glass-box evaluation, but this type of evaluation is more informative with respect to error analysis or future developments of a system.

Automatic procedures can be defined to evaluate an NLP system by comparing its output with the gold standard. Although the cost of producing the gold standard can be quite high, automatic evaluation can be repeated as often as needed without much additional costs. However, for many NLP problems, the definition of a gold standard is a complex task. Manual evaluation is performed by human judges, which are instructed to estimate the quality of a system, based on a number of criteria. Although, human judges can be considered as the reference for a number of language processing tasks, there is also considerable variation across their ratings. This is why automatic evaluation is sometimes referred to as objective evaluation, while the human kind appears to be more subjective.

3. CDAC's Initiatives in S & NLP

C-DAC has been a pioneer in developing and proliferating the use of Indian languages on computers. This technology is now extended to include multimedia and multilingual computing solutions covering a wide range of applications such as publishing and printing, word processing, office application suites with language interfaces for popular third party softwares on various operating platforms, electronic mail, machine translation, language learning, video and television and multimedia content in Indian languages. These have been successfully commercialized. Many a times all these activities are intermingled with each other. Therefore, often developing a system in NLP means to either start working from scratch or use the utilities/subprograms from open source with appropriate modifications. Because of this limitation, the speed of development is slow. To handle this Consortia mode was suggested by CDAC to Department of Information Technology under the National Road Map developed for Technology development in Indian Languages. It was suggested that instead of individual Institutes working in isolation, all willing Institutes could be grouped together in forming Consortia to work for common goal and handle Mission mode Projects favoring common mass. Two types of goals were worked out, Short Term and Long Term. In short term goal, smaller applications such as Desktop publishing

were to be developed for all 22 ILs and make them available to common masses free of charge. This Role out Project was funded by Department of Information Technology and CDAC Pune was asked to lead the consortia. CDAC centers at Noida, Banglore, Mumbai and Mohali were involved with them in development. Noida center developed Portal for the proliferation of the products to masses and understand the need of the users to support further development. In Long term Goals 5 mission mode projects were proposed in E-ILMT, IL-ILMT,OCR/OHR, CLIA and Human resource development in Language Technology. CDAC Noida in participating in all five projects and successfully completed tasks defined in 1st phase. 2nd phase is to start in mid of 2010.

3.1 Universal Speech Translation Advanced Research (USTAR) Consortium project:
The objective of the project is to initiate speech to speech translation service among Asian languages. This project aims at 1) Establishment of an international research collaboration group 2) Building large scale speech and language corpora and technologies 3) Initiate speech translation trial service in Asia
The other participating organizations are - ATR (Japan, coordinator), NLPR(China), ETRI(Korea), BPPT(Indonesia), NECTEC(Thailand), IOIT (Vietnam), CDAC(India), National Taiwan Univ. (Chinese Taipei) and A-STAR (Singapore).
The tasks involved are standardization of corpora, linguistic tags information, communication protocols of modules and interface formats for three major modules of Automatic Speech Recognition (ASR), Machine Translation (MT) and Text to Speech Synthesis (TTS). Successful demo of technology to showcase the proof of concept for enabling online multi-lingual and multi-national S2S translation and communication was carried out among various countries participating in the consortia.

3.2 Digital Library
CDAC Noida has been declared as Mega Centre for Digital Library. We have been working in the area since last 8 years and have acquired requisite expertise to digitize contents containing books to age old manuscripts. The data amounting to the tune of 24 million pages of English, Hindi, Arabic, Persian and Urdu have been digitized till date and data has been made available through Digital Library of India portal. All of the data is linked with metadata to enable easy access.

3.3 Cross Lingual Information Access (CLIA)
As the amount of textual data on the Internet increases, there is also an increasing number of people who want to retrieve information in their native language. Many users also have multilingual capabilities that allow them to understand more than one language. This is one of the main reasons behind developing cross-language information retrieval systems (CLIR). In these kind of systems, users can give queries in their native language and retrieve documents, whether in the same language as the query is, or relevant documents are found in any other language. Cross-Language Information Access exploits the advantage of multilingual capability of users and expands search bandwidth by providing the content which is available in other language also. Cross-language information retrieval enables users to enter queries in languages they are fluent in, and uses language translation methods to retrieve documents originally written 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.

3.4 English to IL translation based on Anglabharati approach

English to Urdu and English to Punjabi translation system based on Rule Base Anglbharati approach has been developed under this project. The system has about 56000 words lexical resource, about 1834 rules, besides Morphology and Generator Modules. The system has been alpha tested to produce 48 and 52 % meaningful translation by third party. After incorporating observations by third party and appropriate patches, now the system is giving 82 and 34 % meaningful translations in Punjabi and Urdu. In 2010 the system will further be improved by adding Language Module and improved preprocessor. The support of " Raw Example Base " has also been planned to support the system by statistical approach as well.

Following table shows completed projects by S & NLP lab in last few years.

S.No	Project Title
1.	Gyan Nidhi – Multilingual Corpus
2.	Mobile Digital Library-I
3.	Machine Aided Translation
4.	Translation Support System
5.	On-line IT Terminology
6.	On-line Hindi Vishwakosh
7.	Design & Development of Bharatiya Bhasha Kosh
8.	Tagged Hindi Corpora
9.	TDIL website Phase-I
10.	TDIL website Phase-II
11.	Development and maintenance of TDIL portal and LTDF in DIT
12.	Ministry of Health & Family Welfare
13.	DGH (Ministry of Petroleum & Natural Gas) Phase-I
14.	DGH (Ministry of Petroleum & Natural Gas) Phase-II
15.	Development of SRS for AO/AI
16.	Development of CBT for CBSE
17.	Development of MIS for MMTC
18.	Integrated Word Processor with OCR
19.	Mobile Digital Library (Dware Dware Gyan Sampada)
20.	Test Bed for Translation Support System
21.	Digital Library project for Uttranchal State Government (Pant Nagar & Nainital)
22.	Digital Archiving for preservation of rare manuscripts and old magazines available with Nagri Pracharini Sabha, Varanasi
23.	Creation of Digital Library of books in President House
24.	Development and maintenance of TDIL portal and Language Technology Demonstration Facility in DIT Phase-IV
25.	E-bank bilingual solution for Stellar Informatics

S.No	Project Title
26.	Development of Annotated Speech Corpora for 3 Indian Languages Hindi, Marathi & Punjabi
27.	Speech Corpora development for ELDA France
28.	Speech Corpora development for SAG
29.	Portal development for Vigyan Prasar
30.	Mega Centre- LTCD
31.	TDIL Data Center
32.	Rajbhasha Information Technology Application Promotion Program. RITAP
33.	Web Internationalization Initiative for Indian Languages

In 2009 and 2010 following projects are ongoing.

Sr.No	Project Name
1	IL-IL Machine Translation (RFP)
2	Cross Language Information Access (RFP)
3	OCR
4	Specialized Post Graduate Level Manpower Development for Localization
5	Masters Programme in Computational Linguistics
6	AnglaBharti Machine Translation (RFP)
7	A Star / Ustar
8	Digital Library

Some significant results achieved in the year 2009 -10 are:

- English to Punjabi translation efficiency tested is above 80 %
- English to Urdu translation efficiency tested is above 82 %
- Punjabi to Hindi translation efficiency tested is above 90 %
- Hindi to Punjabi translation efficiency tested is above 74 %
- English to Punjabi and Urdu translation web version is ready for providing services
- Hindi to Punjabi and reverse translation system is ready and has been noted as the Best amongst all translation systems for Indian languages.
- CLIA & R for Punjabi is ready and under testing.
- Integration of OCRs in Indian Languages has been successfully completed and is appreciated by all consortia and testing team members.
- Two M.Tech. Thesis have been completed and 4 are ongoing.
- Three papers are published and three are under review at International Level.
- 27 Copy rights are worked out and in the last phase of submission.
- 9 IPRs have been worked out and will be filed shortly.
- One scientist has registered for Ph.D. with IIT Delhi in the area of translation testing.

4. Future Directions

Statistical natural-language processing uses stochastic, probabilistic and statistical methods to resolve some of the difficulties discussed above, especially those which arise because longer sentences are highly ambiguous when processed with realistic grammars, yielding thousands or millions of possible analyses. Methods for disambiguation often involve the use of corpora and Markov models. Statistical NLP comprises all quantitative approaches

to automated language processing, including probabilistic modeling, information theory, and linear algebra.

CDAC Noida has following plans for the future

1. Increase ties ups with R & D institutes researching in S & NLP at National and International level.
2. Evaluate the user requirements.
3. Generate the specifications of the packages, utilities and services expected by the end users
4. Work out plans for converting lab produce to products and services.
5. Follow software standards in development life cycles.
6. Enrich staff by appropriate trainings and encourage them to carry out meaningful research
7. Develop Language independent Engines and lingual resources required to plug in with these engines to convert them into useful produce.
8. Use statistical approaches and methods, and incorporate them with rule base approaches to increase the functional efficiency for Indian Languages.

CDAC as an Institution has multifold self imposed responsibility because of the expectations from the users. These include not only core research but the products as well. So far CDAC has maintained 47 % of Indian market satisfied but as penetration of it is increasing in every walk of the life, demand from NLP producer is also increasing day by day. While Intuitions like IITs and IIITs are working in core research, Units like CDAC are to work as premier bridges between R & D and Industries to struggle for meeting the user requirements. It is expected of CDAC that CDAC brings research from Lab to Land. CDAC, Noida will work coherently with other CDAC units to achieve the expectations of the users.

Bibliography

- www.cdac.in
- www.cdacnoida.in
- www.ildc.gov.in
- www.tdil.mit.gov.in
- www.brainhat.com
- www.signiform.com
- www.clsp.jhu.edu
- www.lti.cs.cmu.edu
- www.dai.ed.ac.uk/groups/nlp/NLP_home_page.html
- www.ai.mit.edu/projects/infolab
- www-a2k.is.tokushima-u.ac.jp
- www.speechtek.com
- www.lec.com
- www.naturalvoices.att.com
- www.cstr.ed.ac.uk/projects/festival
- wwwtios.cs.utwente.nl
- Nlpers.blogspot.com

About author

Mr. Vijendrakumar Narendraprasad Shukla is Scientist 'F' in C-DAC Noida and heads Speech & Natural Language Processing lab in C-DAC Noida. Prior to joining current Organization, he has been a teaching professional in various academic institutions linked with University of Pune, Marathwada University,University of Bombay and I.I.T.Roorkee. Mr.. Shukla started his career in teaching way back in 1982 and has about 26 years of experience. He received his M. Tech. (Computer Science & Technology) from University of Roorkee. He has a number of research publications at International and National level to his credit. He is a member of various professional bodies like IETE, IEEE, ISTE. He was Manager W3C India office. He worked as a Convener of the National Committee on Technology Development in Indian Languages (TDIL) set up by the Department of Information Technology, Government of India. He was conferred the 1st IETE-Brig M L Anand Award for his outstanding contribution in the field of I.T. His research areas include Digital Library, Data Mining, Formal Computing, Software Engineering, Multilingual Computing, & Open Source Technology.