A REVIEW ON THE PROGRESS OF NATURAL LANGUAGE PROCESSING IN INDIA

Cini Kurian
Department of Computer Applications,
Cochin University of Science and Technology, Cochin, Kerala, India

ABSTRACT
Natural Language Processing (NLP), which is a branch of artificial intelligence, includes speech synthesis, Speech recognition, and Machine translation. Natural Language Processing has a wide range of applications in the Indian context. Most of the rural Indian community is unable to make use of the information technology revolution due to the dominance of English. Developments in the Natural processing Technology will offer universal access to information and services for more number of people in their mother language. The benefits that are expected to accrue as a result of widespread use of local language computing in India are discussed in this paper. India being a multilingual country with 22 official languages and about 1650 dialects, the research in Natural Language Processing faces great challenges in India. This paper makes a survey of the progress made in this field and explains the difficulties and challenges faced by the research community in the field of Natural Language Processing in India, and also suggests some practical solutions.

KEYWORDS: *Natural language processing speech processing, Indian languages, NLP tools*.

I. INTRODUCTION

One of the greatest gifts of human beings is the ability to communicate. Language is the basic tool of communication. The basic system of communication is speech. Written communication is only secondary. Since primary communication is most important, Natural Language Processing has to play a prominent role in the modern communication technology. Consequent to the revolution in information technology, the concept of communicating with machines enables more users to enjoy the benefits of information technology, especially, for the visually challenged and for those who cannot write. Once computer could understand the language of the human beings, it would be possible to withdraw keyboards attached to computers. Speaking (communicating) with computer for applications related to information is the ultimate outcome of Natural Language Processing. Recently, NLP has grown significantly due to many reasons like the extensive use of World Wide Web, high computing power of computers, business applications and mobile communication system
Natural Language Technology development has a variety of applications, especially in Indian scenario, where a number of languages are being used. A large number of utility applications could be developed which will benefit the Indian public in many ways. These utilities would help translation to and from Indian languages, Scan and Read Indian language content in physical form, handle database, access internet and communicate with mother tongue[2]. Besides these, text-to-speech and speech-to-text utilities would also help illiterates, visually challenged and handicapped to be a part of IT revolution. Moreover, high end tools like search engine and content creation would help Indian languages to be elevated on an international level.
The common challenges faced by researches in this field are elaborated in section 2. In section 3 prospects of Natural language Processing is given and in section 4 different research issues are explained and section 5 is the conclusion section.

II. COMMON CHALLENGES

The language "understanding" is the primary step of NLP. But this "understanding" has many factors. Understanding a language means knowing the concepts a word or phrase and knowing how to link those concepts together in a meaningful way. Hence natural-language recognition requires extensive knowledge about the languages and the ability to interpret it. This knowledge interpretation requires high support from the other important branch of computer science – The Artificial Intelligence (AI).

The ambiguity and complexity of the language made this pasture a challenging one for the researchers. This is because the same word in a language is pronounced differently by different people at different times. The language depends on the context, state of mind, caste, creed, community, status, color and geographical location. Hence the challenges in connection with of Natural language processing are unimaginable.

Natural Language processing is multidisciplinary area of science. It requires both knowledge and expertise from linguistics, computer scientists and technology experts. It has to be developed in line with Computational Linguistics (CLI), a scientific study of language from a computational perspective. CLI is an interdisciplinary field which draws on linguistic theory (phonology, syntax, semantics, pragmatics) and computer science (artificial intelligence, theory of computation, programming methods), and to a lesser extent, other disciplines such as philosophy, cognitive science, and psychology.

III. PROSPECTS OF NATURAL LANGUAGE PROCESSING IN INDIA

Compared to the international scenario, Natural Language processing in India is in its infancy stage. But the prospects of NLP are more in India compared to other courtiers. This is because Indian community faces a "Digital Divide" due to dominance of English in Information and Communication Technology. If tools for information processing and communication are available in local languages, it can put an end to "the digital divide "and can pave way to the "digital unite and knowledge for all" [2].

The following are the benefits that can be acquired as a result of natural language computing in India.

i) Although literacy rate of India is above 65%, less than 6% of India's total population uses English for communication. Since the World Wide Web becomes common any sort of information is available to common man. Moreover, since it has been accepted and implemented even by local bodies make it more important. Therefore it is imperative that the about 95% of our population cannot enjoy the benefits of this revolution. If these information is available in local languages, India could also be benefited by this technology revolution and could stand along with developed countries. NLP has a vital role to play in this process of information technology revolution. Thus the concept of "Global Village" could never be achieved without NLP. NLP will also accelerate the growth of E-learning in India.

ii) NLP can bridge digital divides not only between the literate and illiterate, but also between the disabled and the rest. In addition to this, physically challenged people could also be brought to the forefront of digital revolution. Moreover, NLP can assist the blind people in meeting their academic objectives. Speech to text and Text to speech tools are some of these categories.

iii) E-Governance is the public sector's use of information and communication technologies with the aim of improving information and service delivery, encouraging citizen participation in the decision-making process and making government more accountable, transparent and effective. Therefore, NLP is having a significant role in the implementation of E-governance, since English continues to be a mode of communication in higher education, judiciary, corporate sector and Public administration.

iv) In a large multi-lingual society like ours, there is a great demand for translation of documents from one language to another. The Union Government's official documents and reports are in bilingual (Hindi/English) whereas most of the state government's work in their respective regional languages. So there arise the need to translate these reports and documents to the respective regional languages. Currently these works are being done by human translators. These may not be perfect and hence not dependable. A machine assisted translation system could increase the efficiency of human translators.

v) As the trade and business are ever growing, the people migrated to do and expand business are forced to learn more than one language to communicate with others. The NLP applications can play a significant role in this area. Rural Traders will be benefited by updated market trends daily. Hence their products could be traded in real time. NLP tools make them aware of the market fluctuations and help them improving their business, thus Economic Prosperity could also be improved.
vi) Development of NLP research activities will open up many employment opportunities in all related areas.
vii) Language research should be promoted so that the old method of studying language would switch to new form and shape.

IV. RESEARCH ISSUES IN INDIA

The barriers of NLP researchers in the Indian context are discussed below. As discussed earlier, unlike other countries, in India, the basic issue is the multiliguality. Moreover the same language is pronounced differently in different locations. Words are also having different meanings at different locations. This makes the researchers work more complex and intricate.
The other constraints are:

4.1 Lack of Annotated Corpora

In modern linguistic, Corpus is the machine readable form of the large collection of structured text in written or spoken form [1]. If corpora can give some linguistic information it is called annotated Corpora. Corpus development gained much attention due to recent statistics based natural language processing. It has new applications in Language Technology, linguistic research, language education, information exchange etc. Corpus based Language research has an innovative outlook which will discard the aged linguistic theories. Large collection of corpora is used for training which is an important factor in AI based systems [4].
In English and in other languages many path breaking researches have been done and many pioneering computer based systems have been developed using language corpora. Importance of language corpora is recognized in many countries. However, as far as India is concerned, using corpora in language and NLP research is a time taking process as it is difficult to capture fancy of Indian linguistics because of its diversity. India, as a multilingual country realized its prospects and DOE (Department of Electronics, Govt. of India) under the TDIL (Technology Development for Indian Languages) program has initiated some work on corpora development of all major Indian languages [5]. Under this program, in association with CIIL (Central Institute of Indian Languages) , machine readable corpora for major languages has been developed . While comparing with British National Corpus (BNC) which contains data obtained from people on all walks of life, we are in the infancy stage. Speech corpora is also in its primary phase in India. About 50 hours of annotated speech corpora for Hindi, Marati, Punjabi, Bengali, Assamese, and Manipuri languages have been developed by C-DAC (Centre for Development of Advanced Computing) Noida and C-DAC Calcutta. Corpus generation in India is facing several problems due to lack of a centralized authority (a consortium). Many organizations and institutes have collected corpus for their own research activities. However, these resources are not available to all groups of people working for corpus generation. TDIL and CIIL have taken some initiatives and put the data on the web and their contributions have been appreciated. But there should be national archive for Indian language corpora, so that all corpora will be systematically preserved, documented, distributed, accessed by the users. The Linguistic Data Consortium (LDC), European Language Resources Association (ELRA), and The International Computer Archive of Modern and Medieval English (ICAME) are the good models for this.

4.2 Lack of NLP Tools

There is an acute scarcity of online lexical resources for Indian languages. Building a Natural Language Processing System without basic lexical resources is almost impossible. Those who needs to start building an NLP system, has to start from the scratch with respect to NLP tools like corpora, lexicons, taggers, dictionaries, morphological generator, POS(Part Of Speech) tagger etc. This is a great challenge for researchers in India. This is a Herculean task which cannot be done by a single

group. Whatever little exists has been developed for specific groups and cannot be shared easily. Sharing of resources is the means to help NLP projects to take off swiftly.
Fortunately, the Ministry of Information Technology is taking concrete steps towards creating these resources. Funded by Ministry of Information Technology, IIT (Indian Institute of Technology) Bombay has built Hindi Wordnet and Marati Wordnet [6] . This will help and facilitate NLP Research and Development in the country. Linguistic Data consortium for Indian languages (LDC-IL) by CIIL, Mysore and the open source initiative LERIL(Lexical Resource for Indian languages) are other major initiatives in this sector[7].

4.3 Lack of Standards

There is an urgent need to popularize standards for the following levels: Script level , Font level, Access level (indexing, sorting, and metadata) and Input level (input/keyboard standards). Moreover, transliteration rules are also to be standardized. Although, some standard drafts have been made and presented, such as the 8-bit ISCII (Indian Script Code for Information Interchange) or 16-bit Unicode for script standardization, ISFOC (Intelligence Based Script Font Code) for fonts, and INSCRIPT(Indian Script) phonetic keyboard layout, the final standards have not yet been suggested and fixed[2,3].
Standards are necessary for any type of technology. In the absence of popular standards, developers will work for solutions based on the proprietary technology. This will result into isolated solutions that freeze sharing of software, data and fonts. The proprietary nature of encoding restricts complete dependence on a single developer and for the post stages of solutions usage of such data is not dependable. Unicode consortium is working on for finding better solutions and players like IBM, Red Hat, Microsoft, Oracle are also working towards this direction. Hopefully better solutions are expected in the near future.
India has to actively take part in international initiatives to remain in sync with the global developments on standards to ensure building of world–class solutions in local language. Representations in global initiatives such as Text Encoding Initiative (TEI) and Open Language Archives Community (OLAC) is needed for India to address its unique language requirements. This will ensure availability of global contents in Indian languages and it will benefit the local language information-seekers[8].

4.4 Lack of Interaction Between Linguistics and Computer Programmers

Any technology development leading to formulation of an end user product, should be continuously evaluated and monitored.. This should be benefited to different groups working on the same issues to take alternative approaches. The best approach should be taken and further developed so that the end product should be ideal. This approach is being adopted in advanced countries to make rapid development in technology. In the case of NLP in India similar evaluation process is of urgent need, as people working on this field is unaware of the best approaches to be taken and hence confused. An evaluation committee should identify important approaches and recommend the best approaches so that they can be integrated into end-user products.. The acceptability of any product is decided by the end users. To prohibit the use of pirated products, there should be a certification authority for NLP products. Unfortunately, there is no such certification authority in India.

4.5 Lack of Evaluation and Certification

Any technology development leading to formulation of an end user product, should be continuously evaluated and monitored.. This should be benefited to different groups working on the same issues to take alternative approaches. The best approach should be taken and further developed so that the end product should be ideal. This approach is being adopted in advanced countries to make rapid development in technology. In the case of NLP in India similar evaluation process is of urgent need, as people working on this field is unaware of the best approaches to be taken and hence confused. An evaluation committee should identify important approaches and recommend the best approaches so that they can be integrated into end-user products.. The acceptability of any product is decided by the end users. To prohibit the use of pirated products, there should be a certification authority for NLP products. Unfortunately, there is no such certification authority in India.

4.6 Lack of Consolidated Efforts

Fragmentation of research activities is another major challenge in India. From the literature survey of NLP in India , it is clear that there are so many ongoing activities at different institutions and research organizations, but they have not come up with a good products. This is because of the lack of public availability of the primary resources. So many resources developed in utilizing the project funds are not available in the public domain. These works have been done repeatedly wasting time and money. A consortium needs to be set up which can distribute resources for researchers, World's most success data consortium (Linguistic Data Consortium) at University of Pennysula is the best model on this [8]. Resources and data developed by different research organizations should contributed to the consortium and industries and researches can take the data for research from the consortium.

In India, under the National –Roll-Out project [9], C-DAC , Pune has made efforts to consolidate the technological developments. The project was successful in developing Basic Information Processing Kits are made available in CD's in major Indian languages.

4.7 Lack of Real Learning from Other's Work

Learning from published works is a crucial feature in any research area. New work has to be started from where the predecessor has completed. Research and Development (R & D) in NLP can be boosted up only if new techniques and technology is continued by learning the problems faced by the predecessor. The recent two main machine translation projects funded by Government of India experienced similar problem [10, 11].

The two problems mentioned above could be resolved to a greater extent by forming an Association of researchers and by publishing updates of developments on NLP. Moreover such an association should make available a platform for knowledge sharing. All those involved in this area should be members of the Association.. In India there is an NLP Association, Natural Language Processing Association (NLPAI) constituted in 2002 which is conducting Conferences on Natural language processing [12]. But the representation in the association is very low. Vishwabharat is the only journal published by TDIL in this line.

4.8 Lack of Education and Training Institutes

Growth of any technology depends on the proper education and training. Once a new technology is introduced, it should be made common and accessible to public, to drive the technology forward. Hence it is to be included in the academic curriculum. Since NLP development is a major requirement for E-governance, sincere efforts from government agencies and departments are to be initiated.

Natural language processing projects requires coupling of efforts from linguists, computer scientists and language technology experts. Hence the curriculum should be designed for such an interdisciplinary course. and should be introduced at least in M.Phil or M.Tech level

In India, NLP based course or training is very odd, except some short term course conducted by C-DAC Trivandrum and Indian Institute of Information Technology, Hyderabad (IIIT). But this is too minimal compared to the research centers in foreign universities like Carnegie Mellon University (CMU).

V. CONCLUSION AND FUTURE WORK

NLP application has tremendous potential in Indian scenario. Although literacy rate of India is above 65%, less than 6% of India's total population uses English for communication. Since the internet has become universal, common man now mainly depend, the same for any sort of information and communication. Therefore it is imperative that the about 95% of our population cannot enjoy the benefits of this internet revolution. If this information is available in local languages, India could also be benefited by this technology revolution and could stand along with developed countries. It would be a vital step in bridging the digital divide between non English speaking people and others. Since there is no standard input for Indian languages, it eliminates the key board mapping of different fonts. In Indian scenario, where there are about 1670 dialects of spoken form, speech recognition technology has wider scope and application. Moreover this would also generate more employment opportunities

directly and indirectly. Although several researches are ongoing, India has to work more to be in par with technologically advanced countries.. India can become global player in NLP if proper encouragement and facilities are provided by the Government and private agencies for research and development activities.

The future work will focus on the development of any specific NLP tool development like text to speech tools, speech to text tool or machine translation tool. Towards this goal, the preliminary work will focus on collection of corpus for the proposed tool.

REFERENCES

[1]. Dash, N S and B B Chaudhuri. "Why do we need to develop corpora in Indian languages", *International Conference on SCALLA, Banglore, 2001*

[2]. Frost & Sullivan. Local Language Information Technology Market in India.*Ministry of Communication and Information Technology* - 2003Gizem, Aksahya & Ayese, Ozcan (2009) *Coomunications & Networks*, Network Books, ABC Publishers.

[3]. Murthy, B K and W R. Despande. Language technology in India: past, present, and the future. *In the Proceedings of the SAARC Conference on extending the use of Multilingual and Multimedia Information Technology (EMMIT'98). Pune, India*

[4]. N Niladri Sekhar Dash. "Language Corpora: Present Indian Need", *International Conference on SCALLA, Nepal, 2004* . R M K Sinha. " Machine Translation : An Indian Perspective " , *Proceedings of the Language Egineering Conference (LEC'02*)

[5]. R M K Sinha. " Machine Translation : An Indian Perspective " , *Proceedings of the Language Engineering Conference (LEC'02*)

[6]. Pandey et al, "From Digital Divide to Digital Opportunity", *Proceedings of IEEE Region10 Conference,* 19-21 Nov. 2008, page(s): 1 -6.

[7]. Vishal Goyal and Gurpreet Singh Lehal. "Web Based Hindi to PunjabiMachine Translation System", *Journal of Emerging Technologies in Web Intelligence*, Vol. 2, No. 2, May 2010, pg(s):148-151.

[8]. Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah and *Shata-Anuvadak: Tackling Multiway Translation of Indian Languages*, LREC 2014, Rekjyavik, Iceland, 26-31 May, 2014

[9]. Raj Dabre, Archana Amberkar and Pushpak Bhattacharyya, *A Way to Break Them All: A Compound Word Analyzer for Marathi*, ICON 2013, Noida, India, 18-20 December, 2013

[10]. Balamurali A.R.and Ritesh Khapra , Lost in Translation: An Empirical Study on Cross Language Sentiment Analysis, 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2013), Samos Island, Greece, 24-31 Mar, 2013

[11]. Pushpak Bhattacharyya, Natural Language Processing: A Perspective from Computation in Presence of Ambiguity, Resource Constraint and Multilinguality, CSI Journal of Computing, Vol. 1, No. 2, 2012

AUTHORS

Cini Kurian was born in Cochin , Kerala , Indian in 1979 .She received the Bachelor degree in Computer Science from Mahatma Gandhi University, Kottayam in 1999 and Masters degree from Shivaji University , Kolhapur in the year 2002. She did M.Phil from Bharatidsan University in the year 2006. She is currently pursuing Ph.D. degree with the Department of Computer Applications, Cochin University of Science and Technology. Her research interests includes natural language processing, Automatic Speech recognition and artificial neural networks.

.