International Journal of Applied Research 2017; 3(6): 1416-1421

ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 5.2
IJAR 2017; 3(6): 1416-1421
www.allresearchjournal.com
Received: 14-05-2017
Accepted: 11-06-2017

Dr. N Murali
SV Vedic University, Tiruapti, Andhra Pradesh, India

Correspondence
Dr. N Murali
S.V. Vedic University, Tiruapti, Andhra Pradesh, India

A brief history of research in natural language processing in Sanskrit

Dr. N Murali

Abstract
Natural Language Processing is a branch of Artificial Intelligence. Research on Natural Language Processing has started in western countries in 1940s itself. But in India the research on Natural Language Processing has started only early in 1990s. Linguistically India is a diverse country. There are more than 1000 languages are being in use in India. It is true that all Indian languages share common properties due to the influence of Sanskrit on them. Keeping this in view some research groups around the globe are working on Sanskrit Computational Linguistics or Natural Language Processing. In this paper an attempt is made to shed the light on the research done in this field till now and this will help the young researchers where to start from.

Keywords: Natural language processing, computational linguistics, artificial intelligence

1. Introduction
Natural Languages (NL) are the languages used by human beings as a medium of communication among themselves. Each and every language in the world has a history of some hundreds of years. In all these years each and every language has undergone many changes during the course of time. Due to foreign trade or tourism and exchange of language and culture has occurred. Some languages have accepted some foreign words, new languages have emerged as a result of cultural transformations and some languages have vanished during the time. Modern natural languages (NL) are developing according to their own laws, in each epoch. The structure and use of NL is based on the assumption that the participants of the conversation share a very similar experience and knowledge, as well as manner of feeling, reasoning and acting. In a communication process three processes can be divided into three sub tasks i.e., analyzing, understanding and synthesizing and these three tasks are called language possessing tasks. Automatic language processing agent must be capable of interacting with humans via language, which includes understanding of humans via speech recognition and natural language understanding and communicating with humans via natural language generation and speech synthesis [3]. When the whole or a part of these processes which helps to automate the language processing tasks using computer, then we call this as Natural Language Processing (NLP). Hence Natural Language Processing (NLP) is a branch of computer Science which deals with the analysis, synthesis and understanding of natural language texts using computers.

The main objective of NLP is to build computational models of NL for its analysis and generation and understanding of NL texts. Therefore NLP is a discipline between computer Science and Linguistics and Logic [1].

When human being is to analyze natural language texts, different sources of knowledge like Language Knowledge (Grammar, Lexicon and real world knowledge), common sense (world knowledge, domain specific knowledge, context and culture knowledge) are used to understand the text fully.

Hence a natural language processing system must be capable of using considerable knowledge about the structure of the language.

To conduct research in NLP for any language, one needs the lexical resources of that particular language like–annotated corpus, electronic dictionaries, WordNet, Thesaurus, VerbNet, FrameNet etc., and tools like–Grammar formalisms, reasoning mechanisms, Morphological Analyzers, Local Word Grouper, Part-of-Speech Taggers, Parsers etc.

2. A Brief History of NLP

The field of NLP or Computational Linguistics (CL) has emerged as a separate branch of science and a large number of research groups are working in this field. A survey by Association of Computational Linguistics (ACL) during 1983 stated that nearly 85 universities are conducting courses in NLP/CL [1]. Again during 2004-2008 a survey [1] has been conducted by *Mary D. Taffet* and *Robert Dale* which was sponsored by ACL has stated that nearly 300 universities [2] are conducting courses in NLP/CL. Fortunately, I became instrumental in introducing a post graduate course in NLP in Rashtriya Sanskrit Vidyapeetha, Tirupati in the 2004. Apart from this several research institutions, corporate giants like Google, Microsoft etc., are actively working in this field.

NLP is relatively new in India when compared to Europe and East Asian countries like Japan. Only in 1990s the Indian researchers have started focusing on this interesting field. The Department of Electronics and Information Technology under Ministry of Communication and Information Technology has initiated a project called Technology Development for Indian Languages [3] (TDIL) with an objective of developing information processing tools like creating Multilingual Corpora, OCR, Text-to-Speech, Machine Translation, Generic Software Tools, Localization and standardization of language technology etc. Thirteen resource centers across the country have been identified and supported to develop Indian Language Technology Solutions for 18 Indian languages like Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Marathi, Malayalam, Manipuri, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telugu and Urdu along with two foreign languages i.e., Chinese and Japanese. In this Internet era, due to abundant information on various fields and multiplicity of languages has necessitated to focus on this field. Hence, software giants like Microsoft, Google, and IBM have also started working on Indian languages.

In the era of globalization, localization of the Indian languages became essential for IT companies and the researchers in India began working on this aspect. No doubt, that the main focus in beginning is on developing data processing tools. The milestones achieved by the Indian researchers in this very interesting journey are presented hereunder.

R. Mahesh K. Sinha, and Putcha Narasimham at IIT Kanpur have designed keyboarding schemes which helped in the development of a universal keyboarding scheme and a unified internal code for information exchange that was applicable to all Indic scripts. Later this has led to the development of ISCII (Indian Standard Code for Information Interchange) and then to Unicode. In 1983, the researchers at IIT Kanpur were succeeded in developing Integrated Devanagari Computer (IDC) terminal and the Graphics and Indian Script Technology (GIST) [38].

In 1985, Rick Briggs has presented a paper titled "Knowledge Representation in Sanskrit and Artificial Intelligence" which attracted a lot of young researchers to focus on Sanskrit and Computers [39].

In 1988, Centre for Development of Advanced Computing (C-DAC), Department of Information Technology, Government of India has acquired IDC and GIST technology. Mohan Tambe and his group at C-DAC has enhanced this technology and released word processing, desktop publishing software tools like iLeap, Leap Office etc. [38].

In 1995, the Department of Information Technology, Government of India has sponsored the project on machine aided translation (MAT) from English to Hindi based on AnglaBharathi technology [38].

In 1995, at IIT Kanpur Dr. Vineet Chaitanya Ji along with Dr. Rajeev Sangal has started the Anusaaraka project for machine translation [38].

The major breakthrough in Indian text processing is the introduction of InScript keyboard. A single unified keyboard layout which is in phonetic order for all official Indian scripts facilitated the users to type in any language easily [38].

In 2006, five consortia have been formed by the Ministry of Information Technology, Government of India. Sanskrit to Hindi Machine Translation system is one among the five. The consortium was mooted to avoid the duplication of efforts and to deliver the results. Under this project, Sanskrit-Hindi Accessor Cum Machine Translator (SHMT) has been developed by University of Hyderabad, Hyderabad. The SHMT has been developed based on Anusaaraka guidelines. SHMT translates the given text in Sanskrit language into Hindi language. Various input methods have been provided in its web interface. SHMT consists of different modules like MA, POS Tagger, Chunker, Parser and Translation Modules.

A list of resource centers have been given in table 1 [4].

Table 1: List of TDIL Resource centers

Name of the Institute	Languages
Indian Institute of Technology, Kanpur	Hindi & Nepali
Indian Institute of Technology, Mumbai	Marathi & Konkani
Indian Institute of Technology, Guwahati	Assamese & Manipuri
Indian Institute of Science, Bangalore	Kannada & Sanskrit
Indian Statistical Institute, kolkata	Bengali
Jawaharlal Nehru University, New Delhi	Japanese, Chinese & Sanskrit (Language Learning Systems)
University of Hyderabad, Hyderabad	Telugu
Anna University, Chennai	Tamil
MS University, Baroda	Gujarati
Utkal University, Department of Computer Science and Application	Oriya
Thapar Institute of Engineering & Tech., Patiala	Punjabi
ERDCI, Thiruvananthapuram	Malayalam
CDAC, Pune	Urdu, Sindhi & Kashmiri

The list of ongoing projects in TDIL has been given in table 2 [4].

Table 2: List of ongoing TDIL Projects

Name of the Project	Resource Centre
English to Indian Languages Machine Translation System	CDAC, Pune
English to Indian Languages Machine Translation (MT) System with Angla-Bharti Technology	IIT Kanpur
Indian Language to Indian Language Machine Translation System	IIIT Hyderabad
Sanskrit-Hindi Machine Translation	University of Hyderabad and JNU, New Delhi
Document Analysis & Recognition System for Indian Languages	IIT Delhi
On-Line Handwriting Recognition	I.I.Sc, Bangalore
Cross Lingual Information Access	IIT, Bombay
Speech Corpora & Technologies	IIIT Chennai
Indian Language Corpora Initiative	JNU, New Delhi

Apart from this, many universities and research centers both in India and abroad have concentrated on developing language analysis tools for various Indian languages especially for Sanskrit. The Universities or institutions which are active now in this field are given below:

Department of Sanskrit Studies, University of Hyderabad
Dr. Amba Kulkarni a leading scientist in MTS in India. The department has developed many tools which are made available to public through their website www.sanskrit.uohyd.ac.in/scl/.

Special Centre for Sanskrit Studies, Jawaharlal Nehru University
The Computational linguistics R&D at special centre for Sanskrit Studies J.N.U., started since 2002 lead by Dr. Girish Nath Jha. Research is done in language technology for Sanskrit and other Indian languages. These tools are available at www.sanskrit.jnu.ac.in/index.jsp.

Center for Indian Language Technology (CFILT), IIT Bombay
Prof. Pushpak Bhattacharya at IIT Bombay (currently he is the director of IIT Patna) who is working in NLP and has developed Word Net for Hindi, Marathi and Sanskrit. These tools are available at www.cfilt.iitb.ac.in.

Rashtriya Sanskrit Vidyapeetha, Tirupati
A special course on NLP has been started in the year 2004. The department of Vyakarana and department of computer science are active in this field.
Dr. B.B. Chaudhuri at Indian Statistical Institute, Prof. Harish Karnick from IIT, Kanpur are other notable scholars who are working in this field
Many foreign institutes of repute are working on Sanskrit and Computers. Dr. Gerard Huet from INRIA, France has developed a number of tools for Sanskrit. All these tools are available in his Sanskrit Heritage Site [4]. Harvard University has also developed morphological analyzer for Sanskrit [5]. In Germany 14 Universities are offering courses on Sanskrit and Indology. Many European countries are also teaching Sanskrit.

3. Machine Translation in India
Salil Badodekar [24], Garje and Kharate [40] has given a good information regarding the NLP and MT research activities in India. The Table 3 presents the list of ongoing machine translation projects in India:

Table 3: List of Machine Translation Systems [40]

Name of the Project	Name of the Researcher (s)
Anusaaraka	Dr. Vineet Chaitanya and Dr. Rajeev Sangal
Punjabi to Hindi MT System	Josa G.S. and Lehal G.S. [10, 12]
Mantra MT (English to Hindi MTS)	Bharati [14]
An English Hindi Translation System	Gore L and Patil N [16]
MAT (English to Kannada)	Murthy K [17]
English-Telugu MTS	Bandyopadhyaya S [19]
Telugu-Tamil MTS	Bandyopadhyaya S [19]
OMTrans (English to Oriya MTS)	Mohanty S, Balaantaray RC [20]
MaTra System (English to Hindi) Domain: News	Anantha Krishnan R, Kavitha M, Hegde JJ, Chandra Sekhar, Ritesh Shah, Swani Bade and Sasikumar M [21-22]
English-Kannada MTS **Domain:** Government Circulars	K Narayana Murthy [23]
Tamil-Hindi Machine-Aided Translation System	Sobha L, Pralayankar P and Kavitha V [24]
Sampark System: Automated Translation among Indian Languages	funded by TDIL [25]
UNL-based English-Hindi MTS	Dave S, Parikh J and Bhattacharyya P [26]
AnglaHindi (English to Indian languages)	RMK Sinha and Jain A [27]
Bengali to Hindi MTS	Chatterji S, Roy D, Sarkar S and Basu A
Lattice Based Lexical Transfer in Bengali Hindi MT Framework	Sanjay Chatterji, Praveen Sonare, Sudeshna Sarkar and Anupam Basu [31]
Anubaad (English to Bengali)	Bandyopadhyay S [32]
Vaasaanubaada (Bengali to Assamese News Texts)	Vijayanand K, Choudhury SI and Ratna P [33]
Shiva and Shakti MTS (English to Indian Languages)	Carneige Mellong University of USA, IIIT Hyderabad and IISc, Bengaluru [18, 28]
Hinglish MTS	RMK Sinha and Thakur [34]

English to {Hindi, Kannada, Tamil} and Kannadato Tamil Language-Pair	Balajapally P, P Pydimarri, M Ganapathiraju, N Balakrishna and R Reddy [28]
The MATREX (MT using Example)	Ankit Kumar, Srivastava, Rejwanul Haque, Sudip Kumar Naskar and Andy Way [35-36]
English to Indian Languages MTS (E-ILMT) (English to Indian languages) **Domain:** Health Care and Tourism	Consortium of Nine Institutions [37].

4. Vedas - Computers

The application of NLP techniques in analyzing Vedas is not yet fully explored. Till now computers are being used for type setting of Vedas, playback of Vedic hymns etc. Some self-motivated traditional Vedic scholars and some institutions are working in this interdisciplinary to bring some tools for learners and Vedic lovers. A brief description of those institutes and the tools developed by them are presented below:

4.1 Indian Heritage Group

When the application of computers in Sanskrit and Veda comes, the fore most person to be remembered is Dr. P. Ramanujam of Heritage Group, C-DAC. He has touched almost all crucial areas in Sanskrit and Vedas. The Indian heritage language and computing at C-DAC, Bengaluru is engaged in design and development of NLU system, Vedic and Manuscript processin tools, Development of Analyitical Tools for large, scientific knowledgebase in Grid Computing Environment [5]. The Heritage Group has developed the following software for Sanskrit and Vedas:

4.1.1 DESIKA

NLU System for Sanskrit: DESIKA is a comprehensive package for generation and analysing Sanskrit words.

4.1.2 Ancient Indian Scientific Knowledge Base (C-VYASA)

It is designed to assist in providing Sanskrit content tools for processing, hyper-linking, references etc.

4.1.3 RgVeda Ratnakara

Includes Samhita, Padapatha, Khila, Sarvanukrani: Sayanabhashya, English Translation for select suktas.

4.1.4 YajurVeda Reader

Includes Samhita, Padapatha, Brahmana (Kathaka, Aranyaka and Ekagni-Kanda). YajurVeda by most popular schemes - Pancasat/ Dasati and Arsheya Krama. Retrieve details like Pancasat/ Dasati, Padapatha.

4.1.5 SamaVeda

Rendering SamaVeda was a challenge for long because of its different accent scheme. Now, keying in of SamaVeda is as easy as entering any other text. Currently, this is available in Grantha Script.

4.1.6 Nirukta

Has Yaka's Nirukta and Nighantu with accents where present. Has links between the two, searchable from Nighantu.

4.1.7 Upanga-s Mimamsa Brahma Sutra

Various Dusana - Shastras are provided as searchable, indexed applications. Exhaustive set of original texts covering various works of the three major systems of Vedanta are included.

4.2 Academy of Sanskrit Research (Melkote)

It is conducting research in Sanskrit Speech Synthesis, Natural Language Processing, Machine Translation, Sanskrit teaching through computer media etc. for the last 8 years. It is conducting research in Natural Language Processing of Sanskrit using Artificial Intelligence, Cryptography etc. It is also collecting scientific information available in Sanskrit texts [6].

4.3 Avadhoota Dattapeetham

While preservation and propagation of oral tradition of Vedas, Avadhoota Dattapeetham is also striving hard for research in Vedas. The researchers have developed software for creating Vedic texts with svaras [7].

4.4 Sri Venkateswara Vedic University

Even though the University has started in 2006, this nascent University has achieved a remarkable milestone in research in Vedas and allied literature. The University has undertaken a project called Analysis of Varnakrama of Krishnayajurveda Taittireeya Sakha. The Vedic seers have given a clear definition of each syllable, its origin, how to pronounce and the length of time it should be pronounced etc. Under this project software was created to analyse each syllable in mantras including svaras covering all aspects of traditional Varnakrama and this may give clues for speech processing [8].

5. Conclusion

Natural Language Processing and its history was briefly described in this paper. The status of NLP in India, the institutes which are currently working in NLP was also mentioned here. The list of resource centres and ongoing research projects are also presented. A brief description of the Machine Translation tools are discussed. Even though the research in NLP is relatively new to Indian languages, our researchers have made a considerable progress in this field and much has to be achieved. Till now, the application of computers in Vedas is limited only for creating texts. Application of NLP techniques in Vedas is not yet started and it requires a high level of understanding which requires further research in NLP in Sanskrit.

6. References

1. Akshar Bharathi, Vineeth Chaitanya, Rajeev Sangal. Natural Language Processing, A Paninian Perspective. Prentice Hall of India. New Delhi; c1996.
2. James Allen. Natural Language Understanding (2nd Ed.). Pearson Education. New Delhi; c2003.
3. Daniel Jurafsky, James H Martin. Speech and Language Processing. Pearson Education. New Delhi; c2004.
4. Murali Nandi. Understanding of Elementary Sanskrit Text at Word Level, Phrase Level and Sentence Level (Unpublished thesis); c2013.
5. www.cdac.in/index.aspx?id=mc_hc_IHLC_resch_devlp
6. www.sanskritacademy.org/research.htm
7. www.vedanidi.in

8. Murali N, Sarma KTRK. Vaidikavarnadharmanam Sanganaka Yantraropanam. In Vedanga Saurabham Proceedings of National Seminar on the importance of the Six Ancillary Disciplines (Shadangas) in preserving the Vedic Learning. Sri Venkateswara Vedic University, Tirupati; c2014. 3rd-5th Jan. p. 135-142. ISBN: 9783981887356
9. Akshar Bharati, Chaitanya Vineet, Amba P Kulkarni, Rajiv Sangal. Anusaaraka: Machie Translation in Stages. Vivek, a quarterly in Artificial Intelligence, NCST, Mumbai. 2001;10(3):22-25.
10. Josan GS, Lehal GS. A Punjabito Hindi Machine Translation System. In Poceedings of COLING-2008: Companion volume: Posters and Demonstrations, Manchester, UK; c2008. p. 157-160.
11. Sugata Sanyal, Rajdeep Borgohain. Machine Translation Systems in India. Cornel University Library; c2013. Arxiv.org/ftp/arxiv/papers/1304/1304.7728.pdf
12. Goyal Vishal, Singh Gurpreet Lehal. Web Based Hindi to Pnjabi Machine Translation System. International Journal of Emerging Technologies in Web Intelligence. 2010;2(2):148-151.
13. Josan Gurpreet Singh, Kaur Jagroop. Punjabi To Hindi Statistical Machine Translations. International Journal of Information Technology and Knowledge Management. 2011;4(2):459-463.
14. Naskar Sudip, Shivaji B. Use of Machine Translation in India: Current Status. AAMT Journal; c2005. p. 25-31.
15. Darabari Hemant. Computer Assisted Translation System: An Indian Perspective. In Proceedings of MT Summit VII, Thailand; c1999.
16. Gore Lata, Patil Nishigandha. English to Hindi - Translation System. In Proceedings of Symposium on Translation Support Systems. IIT Kanpur; c2002. p.178-184
17. Murthy K. MAT: A Machine Assisted Translation System. In Proceedings of Symposium on Translation Support System (STRANS-2002), IIT Kanpur; c2002. p. 134-139.
18. Akshar Bharati, Moona R, Reddy P, Sankar B, Sarma DM, Rajeev Sangal. Machine Translation: The Shakti Approach. Pre-Conference Tutorial, ICON; c2003.
19. Parameswari K, Sreenivasulu NV, Uma Maheswar Rao G, Christopher M. Development of Telugu-Tamil Bidirectional Machine Translation System: A special focus on case divergence. In Proceedings of 11th International Tamil Internet Conference; c2012. p. 180-191.
20. Mohanty S, Balabantaray RC. English to Oriya Translation System (OM Trans); c2004. Cs. pitt.edu/change/cpol/c087.pdf
21. Ananthakrishnan R, Kavitha M, Hegde J Jayaprasad, Sekhar Chandra, Shah Ritesh, Sawani Bade, *et al*. MaTra: A Practical Approach to Fully-Automatic Indicative English-Hindi Machine Translation". In Proceedings of MSPIL-06; c2006.
22. CDAC, Mumbai. MaTra: An English to Hindi Machine Translation System. A report by CDAC Mumbai formerly NCST; c2008.
23. Latha R Nair, David Peter S. Machine Translation Systems for Indian Languages. International Journal of Computer Applications (0975-8887) 2012, 39(1).
24. Badodekar Salil. Translation Resources, Services and Tools for Indian Languages. A report of Centre for Indian Language Technology, IITB; c2004. http://www.cfilt.iitb.ac.in/Translations-survey/survey.pdf
25. Sampartk: Machine Translation System among Indian Languages; c2009. http://sampark.iiit.ac.in
26. Jain Manoj, Damani P Om. English to UNL (Interlingua) Enconversion. In proceedings of 4th Language and Translation Conference (LTC-09); c2009.
27. Sinha RMK, Jain A. Angla Hindi: An English to Hindi Machine-Aided Translation System. International Conference AMTA (Association of Machine Translation in the Americas); c2002.
28. Sanjay Kumar Dwivedi, Pramod Premdas Sukhadeve. Machine Tranlation System In Indian Perspectives. Journal of Computer Science. 2010;6(10):1082-1087. ISSN 1549-3636.
29. projects.uptuwatch.com/cs-it/anubharti-an-hypbrid-example-based-approach-for- machine-aided-trapnslation/
30. Sanjay Chatterji, Devshir Roy, Sudeshna Sarkar, Anupam Basu. A Hybrid Approach for Bengali to Hindi Machine Translation. In proceedings of ICON-2009, 7th International Conference on Natura, Language Processing; c2009. p. 83-91.
31. Sanjay Chatterji, Praveen Sonare, Sudeshna Sarkar, Anupam Basu. Lattice Based Lexical Transfer in Bengali Hindi Machine Translation Framework. In Proceedings of ICON-2011: 9th International Conference on Natural Language Processing, Macmillan Publishers, India; c2011. ltrc.iiit.ac.in/proceedings/ICON-2011.
32. Bandyopadhyay S. ANUBAAD: The Translator from English to Indian Languages. In proceedings of the VIIth State Science and Technology Congress, Calcutta, India; c2004. p. 43-51
33. Vijayanand K, Sirajul Islam Choudhury, Pranab Ratna. Vaasaanubaada - Automatic Machine Translation of Bilingual Bengali-Assamese News Texts. In proceedings of Language Engineering Conference-2002, Hyderabad, India; c2004.
34. Sinha RMK, Anil Thakur. Machine Translation of Bi-lingual Hindi- English (Hinglish) Text. In Proceedings of 10th Machine Translations Summit organized by Asia-Pacific Association for Machine Translation (AAMT), Phuket, Thailand; c2005.
35. Ankit Kumar Srivastava, Rejwanul Haque, Sudip Kumar Naskar, Andy Way. The MATREX (Machine Translation using Example): The DCU Machine Translation System for ICON 2008. In Proceedings of ICON-2008: 6th International Conference on Natural Language Processing; c2008. http://ltrc.iiit.ac.in/proceedings/ICON-2008.
36. Yanjun Ma, John Tinsley, Hany Hassan, Jinhua Du, Andy Way. Exploiting Alignment Techniques in MATREX: the DCU Machine Translation System for IWSLT 2008. In proceedings of IWSLT 2008, Hawaii, USA; c2008.
37. Ananthakrishnan R, Hegde Jayaprasad, Pushpak Bhattacharya, Ritesh Shah, Sasikumar M. Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In

Proceedings of International Joint Conference on NLP (IJCNLP08), Hyderabad, India; c2008.
38. Sinha RMK. A Journey from Indian Scripts Processing to Indian Language Processing. IEEE Annals of the Hisotry of Computing. 2009;31(1):8-31.
39. Briggs R. Knowledge Representation in Sanskrit and Artificial Intelligence. AI magazine. 1985;6(1):32-39.
40. Garje GV, Kharate GK. Survey of Machine Translation Systems In India. International Journal on Natural Language Computing (IJNLC). 2013;2(4):47-67. DOI: 10.5121/ijnlc.2013.2504