LANGUAGE IN INDIA
Strength for Today and Bright Hope for Tomorrow
Volume 12 : 4 April 2012
ISSN 1930-2940

Language Technology in India

Keshav Niranjan, Ph.D. Candidate

==

Introduction

In a manner of speaking, we may say that India is a multilingual country where the spoken language changes after every 50 kilometers. Therefore there is no single universal language. Indian languages are classified into five languages families Indo-Aryan (76.87 % speakers), Dravidian (20.82 % speakers), Austro-Asiatic (1.11 % speakers), Tibeto-Burman (1% speakers) and Andmanese (0 % speakers) (Jha, 2003) [1,2]. Constitution of India recognizes 24 languages and 12 scripts. After the information technology revolution which requires the knowledge of English for the manipulation of digital data, information and knowledge store in database and manipulation of the database, i.e., Structured Query Language (SQL), demand for the acquisition of English has increased. In India only 5% of the population is familiar with English and the rest of the population uses different languages. Nearly 50% of Indian population speaks Hindi but it is largely untouched by developments in modern technical knowledge resource. Those who have no knowledge of English may have difficulty in using the knowledge resource. They are also not in a position to enrich the resource due to unfamiliarity with English. Therefore, lack of knowledge of English becomes a hindrance in the progress of science and technology in India due to language barrier.

Government Effort for Evolving Language Technology

Indian government was aware about this fact. Since 1970, the Department of Electronics and the Department of Official Language were involved in developing the Indian language Technology. Consequently ISCII (Indian Script Code for Information Interchange) is developed for Indian languages [3] on the pattern of ASCII (American Standard Code for Information Interchange). Also "**I**ndian languages **Trans**literation" (**ITRANS**) developed by Avinash Chopde [4] and ITRANS represents Indian language alphabets in terms of ASCII (Madhavi et al, 2005) [5]. The Department of Information Technology under Ministry of Communication and Information Technology is also putting the efforts for proliferation of Language Technology in India, And other Indian government ministries, departments and agencies such as the Ministry of Human Resource, DRDO (Defense Research and Development Organization), Department of Atomic Energy, All India Council of Technical Education, UGC (Union Grants Commission) are also involved directly and indirectly in research and development of Language Technology. All these agencies help develop important areas of research and provide funds for research to development agencies [6]. As an end-result IndoWordNet was developed for the Indian languages on the pattern of English WordNet [7,8].

TDIL Program

Government of India launched TDIL (Technology Development for Indian Language) program. TDIL decides the major and minor goal for Indian Language Technology and provide the standard for language technology [9] TDIL journal ***Vishvabharata*** **(Jan 2010)[10]** outlined short-term, intermediate, and long-term goals for developing Language Technology in India.

Development of Language Corpora in Indian Languages

Kolhapur Corpus of Indian English (**KCIE**) was the first Indian language corpora for Indian English, which was developed under the leadership of Prof. S.V. Shastri at the Shivaji University, Kolhapur, India in 1988. **KCIE** contains approximately one million words of Indian English drawn from materials published in the year 1978. This is collected for a comparative study among the American, the British, and the Indian English (Dash) [11]. Central Institute of Indian Language **(CIIL)** is a nodal agency for development of Indian Language Corpora. It has co-coordinated with various Indian agencies and Universities for developing more than 45 million corpora in Scheduled Language of India which is also a part of **TDIL** programme [12]. **Enabling Minority Language Engineering** (**EMILLE)** program provides the corpora, architecture and tool for Asian languages. It has a monolingual corpus which contains approximate 96,157,000 words and a parallel corpus consists of 200,000 words of text in English which helps in the translation of Bengali, Hindi, Punjabi and others languages.

C-DAC Noida has developed the parallel text corpus ***Gyan-Nidhi*** for 12 Indian languages (Hindi, Punjabi, Gujarati, Marathi, Tamil, Telugu, Kannada, Nepali, Oriya, Malayalam, Bangla, Assamese) and English. **Gyan-Nidhi** is also a multilingual parallel corpus, which is a repository of 'One Million Pages' of knowledge based text [13].

Mahatma Gandhi International University has started the project ‘Hindi Samghraha’ for repository of Hindi words database and dialect mapping of Hindi [14]. Department of Information Technology of Government of India has started the project for developing the Indian language Corpora, Indian Language Corpora Initiative (**ILCI**). **ILCI** is a consortium project for building the parallel annotated corpora under the leadership of Dr. Girish Nath Jha, JNU, New Delhi. It involves 11 Indian languages and also English [15].

Machine Translation in India

Although Translation in India is old, Machine Translation is comparatively young. Earlier efforts in this field have been noticed since 1980, involving different prominent Institutions such as **IIT** Kanpur, **University of Hyderabad, NCST** Mumbai and **CDAC** Pune. During late 1990 many new projects initiated by **IIT** Mumbai, **IIIT** Hyderabad, **AU-KBC** Centre, Chennai and **Jadavpur University,** Kolkata [16] were undertaken. **TDIL** has started a consortium mode project since April 2008, for building computational tools and Sanskrit-Hindi MT under the leadership of Amba Kulkarni (University of Hyderabad). The goal of this Project is to build children’s stories using multimedia and e-learning content.

Anglabharati

IIT Kanpur has developed the Anglabharti Machine Translator technology from English to Indian languages under the leadership of Prof. R.M.K Sinha. It is a rule-based system and has approximately 1750 rules, 54000 lexical words divided into 46 to 58 paradigm [17]. It uses pseudo Interlingua named as PLIL (Pseudo Lingua for Indian Language) as an intermediate language. The architecture of **Anglabharti** has six modules: Morphological analyzer, Parser, Pseudo code generator, Sense disambiguator, Target text generators, and Post-editor [18]. Hindi version of Anglabharti is **AnglaHindi** which is web based application which is also available for use at http://anglahindi.iitk.ac.in. To develop automated translator system for regional languages, Anglabharti architecture has been adopted by various Indian institutes for example, IIT Guwahati.

Anubharti

Prof. R.M.K. Sinha developed **Anubharti** during 1995 at IIT Kanpur. **Anubharti is** based on hybridized example-based approach. The Second phase of both the projects (Anglabharti II and Anubharti II) has started from 2004 with new approaches and some structural changes.

Anusaaraka

Anusaaraka is a Natural Language Processing (NLP) Research and Development project for Indian languages and English undertaken by CIF (Chinmaya International Foundation). It is fully-automatic general-purpose high-quality machine translation systems (FGH-MT) [20]. It has software which can translate the text of any Indian language(s) into another Indian Language(s), based on Panini Ashtadhyayi (Grammar rules).It is developed at the International Institute of Information Technology, Hyderabad (IIIT-H) and Department of Sanskrit Studies, University of Hyderabad.

MaTra

MaTra is a fully-automatic indicative English to Hindi translation system. It supports all kind of domains but gives more accurate results in especially ‘News’ and ‘Medical’ domains.

Mantra

Machine Assisted Translation Tool (**Mantra**) is a brain child of Indian Government during 1996 for translation of Government orders, notifications, circulars and legal documents from English to Hindi. The main goal was to provide the translation tools to government agencies. Mantra software is available in all forms such as desktop, network and web based [21]. It is based on **Lexicalized Tree Adjoining Grammar (LTAG)** formalism to represent the English as well as the Hindi grammar [22]. Initially, it was domain specific such as Personal Administration, specifically Gazette Notifications, Office Orders, Office Memorandums and Circulars, gradually the domains were expanded. At present, it also covers domains like Banking, Transportation and Agriculture etc. Earlier Mantra technology was only for English to Hindi translation but currently it is also available for English to other Indian Languages such as Gujarati, Bengali and Telugu. **MANTRA-Rajyasabha** is a system for translating the parliament proceedings such as papers to be laid on the Table [PLOT], Bulletin Part-I, Bulletin Part-II, List of Business [LOB] and Synopsis[23]. Rajya Sabha Secretariat of Rajya Sabha (the upper house of the Parliament of India) provides funds for updating the **MANTRA-Rajyasabha** system.

UNL-based MT System between English Hindi and Marathi

IIT Bombay has developed the **Universal Networking Language (UNL)** based machine translation system for English to Hindi Language**. UNL** is United Nations project for developing the Interlingua for world’s languages. **UNL**-based machine translation is developing under the leadership of Prof. Pushpak Bhattacharya IIT Bombay.

English-Kannada MT System

Department of Computer and Information Sciences of Hyderabad University has developed an **English-Kannada MT system**. It is based on the transfer approach and Universal Clause Structure Grammar (UCSG).This project is funded by the Karnataka Government and it is applicable in the domain of government circulars.

SHIVA and SHAKTI MT

Shiva is an Example-based system. It provides the feedback facility to the user. Therefore if the user is not satisfied with the system generated translated sentence, then the user can provide the feedback of new words, phrases and sentences to the system and can obtain the newly interpretive translated sentence. Shiva MT system is available at (http://ebmt.serc.iisc.ernet.in/mt/login.html).

Shakti is a statistical approach based rule-based system. It is used for the translation of English to Indian languages (Hindi, Marathi and Telugu). Users can access the **Shakti** MT system at (http://shakti.iiit.net).[24]

Tamil-Hindi MAT System

K B Chandrasekhar Research Centre of Anna University, Chennai has developed the machine-aided Tamil to Hindi translation system. The translation system is based on ***Anusaaraka*** Machine Translation System and follows lexicon translation approach. It also has small sets of transfer rules [25]. Users can access the system at http://www.au-kbc.org/research_areas/nlp/demo/mat/.

Anubadok

Anubadok is a software system for machine translation from English to Bengali. It is developed in Perl programming language which supports processing of Unicode encoded and text for text manipulations. The system uses the Penn Treebank annotation system for part-of-speech tagging. It translates the English sentence into Unicode based Bengali text. Users can access the system at http://bengalinux.sourceforge.net/cgi-bin/anubadok/index.pl.

Punjabi to Hindi Machine Translation System

During 2007, Josan and Lehal at the Punjab University, Patiala, designed Punjabi to Hindi machine translation system. The system is built on the paradigm of foreign machine translation system such as RUSLAN and CESILKO [26]. The system architecture consists of three processing modules Pre Processing, Translation Engine and Post Processing [27].

Contribution of Private Companies in Evolving the ILT –

Indian language Search Engine Guruji

guruji.com is the first Indian language search engine founded by the two IIT Delhi graduate Anurag Dod and Gaurav Mishra, assisted by the Sequoia Capital. guruji.com uses crawls technology, based on propriety algorithms. For any query, it goes into Indian languages contents deep and tries to return the appropriate output. guruji search engine covers a range of specific content news, entertainment, travel, astrology, literature, business, education and more [28].

Google

Internet Searching giant Google also supports major Indian Languages such as Hindi, Bengali , Telugu ,Marathi ,Tamil ,Gujarati ,Kannada, Malayalam, and Punjabi and also provides the automated translation facility from English to Indian Languages. Google Transliteration Input Method Editor is currently available for different languages such as Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Punjabi, Tamil, Telugu and Urdu.

Microsoft Indic Input Tool

Microsoft has developed the Indic Input Tool for Indianisation of computer applications. The tool supports major Indian languages such as Bengali, Hindi, Kannada, Malayalam, Tamil and Telugu. It is based on a syllable-based conversion model. WikiBhasa is Microsoft multilingual content creation tool for translating Wikipedia pages into multilingual pages. So, source language in WikiBhasa will be English and Target language can be any Indian local language(s).

Webdunia

Webdunia is an important private player which assists the development of Indian language technology in different areas such as text translation, software Localization, and Website localizations. It is also involved in research and development of Corpus creation/collection, and Content Syndication. Moreover, it provides the facility of language consultancy. It has developed various applications in Indian Languages such as My Webdunia, Searching, Language Portals, 24 Dunia, Games, Dosti, Mail, Greetings, Classifieds, Quiz, Quest, Calendar etc. [29].

Modular InfoTech

Modular InfoTech Pvt. Ltd. is a pioneer private company for development of Indian Languages software. It provides the Indian language enablement technology to many state governments and central government in e-governance programs. It has developed the software for multilingual content creation for publishing newspapers and also has developed the qualitative Unicode based Fonts for major Indian languages. It has specifically developed the Shree-Lipi Gurjrati pacakage for the Gujarati language which is useful in DTP sector, corporate offices and e-Governance program of the Government of Gujarat [30].

Conclusion

In this digital era, knowledge or information is created in English. In India a large section (approximately 95% population) is not speaking English and therefore, they are not benefitting from science and technology knowledge. Hence, new efforts and initiatives have been started for evolving language technology in India. Government agencies and private organizations should work collaboratively to develop new areas in language technology and should develop robust technology to benefit the vulnerable sections. In addition to this, Government should also sanction grants for research and development and also develop a special centre for training in language technology

==

Acknowledgement

The author of this paper thanks Dr. Girish Nath Jha and Pinkey Nainwani of Jawaharlal Nehru University (JNU), New Delhi for reviewing and giving their invaluable comments for this article which helped to revise the draft of this article.

===

References

[1] Jha, Girish Nath, 2003, Current *Trends in Indian languages Technology*, Language in India, Volume 3:12

[2] Jha, Girish Nath,India's language diversity and resources of the future: challenges and opportunities, Special Center for Sanskrit Studies, Jawaharlal Nehru University, New Delhi

[3] http://tdil.mit.gov.in/Standards/ISCII.aspx

[4] http://en.wikipedia.org/wiki/ITRANS

[5] Madhavi, G., Balakrishnan, M., Balakrishnan, N., Reddy, R.: Om: One tool for many (Indian) Languages. J. Zhejiang University Science 6A(11), 1348–1353 (2005),

[6] http://tera-3.ul.cs.cmu.edu/conference/2005/16.pdf

[7] Gupta, B. M., Kshitij ,Avinash ,and Verma Charu ,2011, *Mapping of Indian computer science research output*,1999–2008,Scientometrics (2011) 86:261–283 DOI 10.1007/s11192-010-0272-y

[8] English WordNet is collection of thesaurus of English words, which consist of words, semantic synonyms. Developed by George Miller and Christiane Fellbaum at Princeton University in 1985.

[9] http://tdil.mit.gov.in/AboutUs.aspx

[10] http://tdil.mit.gov.in/Publications/Vishvabharat.aspx

[11] http://www.elda.org/en/proj/scalla/SCALLA2004/dash.pdf

[12] http://www.ciilcorpora.net/index.html

[13] http://tdil-dc.in/tdildcMain/articles/644532CORPORA.PDF

[14]http://en.wikipedia.org/wiki/Mahatma_Gandhi_Antarrashtriya_Hindi_Vishwavidyalaya

[15] http://sanskrit.jnu.ac.in/ilci/index.jsp

[16] Rao ,Durgesh, *Machine Translation in India: A Brief Survey* ,National Centre for Software Technology Gulmohar Road 9, Juhu, Mumbai 400049, India.

[17]http://www.cdacnoida.in/ASCNT-2010/Quality%20Improvement/Paper/Testing%20of%20AnglaBharati%20System.pdf

[18] Balyan, Renu, Shukla,V.N. , Unified Lexical Resource for Indian Languages- Based on AnglaBharti Approach

[19] Naskar, Sudip and Bandyopadhyay, Sivaji, *Use of Machine Translation in India: Current Status*, Computer Science and Engineering Department Jadavpur University, Kolkata, India, 700032

[20] Bharati,Akshar , Chaitanya,Vineet , Kulkarni Amba P., Sangal,Rajeev , *Anusaaraka: Machine Translation In Stages*, Vivek, Vol.10, No.3 (July 1997), NCST, Mumbai, pp.22-25.

[21] http://www.cdacindia.com/html/about/success/Mantra.aspx

[22] http://en.wikipedia.org/wiki/MANTRA-Rajbhasha

[23] Naskar, Sudip and Bandyopadhyay, Sivaji, *Use of Machine Translation in India: Current Status*, Computer Science and Engineering Department Jadavpur University, Kolkata, India, 700032.

[24] http://www.cse.iitk.ac.in/chair/lec/2004.09.15.html

[25] Dwivedi ,Sanjay Kumar and Sukhadeve,Pramod Premdas ,2010,*Machine Translation System in Indian Perspectives*, Journal of Computer Science 6 (10): 1111-1116, 2010, ISSN 1549-3636,© 2010 Science Publications

[26] Goyal, Vishal and Lehal, Gurpreet SINGH ,2009, *Evaluation of Hindi to Punjabi Machine Translation System*, IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2009ISSN (Online): 1694-0784,ISSN (Print): 1694-0814

[27] http://aclweb.org/anthology/C/C08/C08-3004.pdf

[28] http://www.india-reports.com/corporate/Guruji.aspx

[29] http://en.wikipedia.org/wiki/Webdunia.com

[30]http://www.modular-infotech.com/html/news.html

==

Keshav Niranjan, Ph.D. Candidate
Singhania University
Pacheribari 333315
Rajasthan
India
keshav.niranjan@gmail.com