Application of Machine Translation Systems in India in Overcoming Digital Divide in Higher Education: A Critical Review

[1*]Dr. Vijay Srinath Kanchi[1], Dr. JagdishKulkarni[2] and Dr. Sudhir S. Patil[3]

[1]Moolji Jaitha College,Jalgaon -425002, Maharashtra, India
[2]Sri RamanandTirthMarathwada University, Nanded
[3]SSBT's College of Engineering and Technology, Bambhori, Jalgaon,

[1]vskanchi@gmail.com, [2]jnkul72@gmail.com, [3]sudhir78_patil@yahoo.com

Abstract

Since the advent of Information and Communication Technologies post 1990, the information resources available to a researcher, student or a teacher have increased manifold times. Now the printed books no longer are at thecentre stage and traditional libraries have certainly lost the sheen. The world is fast transforming into an information society where information rules every aspect of the human life. There are a plethora of information resources available online and e-books and e-journals provide latest trends in a particular field in a much faster way. There are several initiatives on the part of the Government of India to make available e-resources to the end users of higher education through UGC Infonet, Nlist and other portals. Millions of rupees are spent to bring the e resources to the forefront of users. However,since most of the e-resources are published in English language their usage is very dismal in many of the institutes of higher education. Unless a mechanism is developed to provide the e-content produced in English language in Indian languages, their usage would remain paltry and meager resulting in the public spending being tantamount to squandering. Efforts made in this direction by various agencies fall short in many respects. This paper examines the core problem, evaluates various efforts in this direction and suggests some remedial measures.

***Keywords**: e-resources, automatic translation, Machine translation, digital divide, language barrier.*

1. Introduction

The Information and Communication Technologies that made rapid inroads post 1990, had left no aspect of human life untouched and their impact is also very evident on the educational sector as well. Today libraries are shifting their role from the custodians of traditional information resources to the providers of customized solutions to the specific information needs of the users by expanding their horizons to e-resources as well. Quadri (2012) rightly points out that 'widespread use of computers, increased reliance on computer networks, rapid growth of the Internet and explosion in the quality and quantity of information compelled libraries to adopt new means and methods for the storage, retrieval and dissemination of information'. These information and communication technologies have made locating, retrieving, storing and disseminating information very efficient, making vast information resources readily accessible at the click of a mouse. While a conventional and traditional library always served its clientele within the constraints of time and space, the digital resources are enabling the libraries to serve their

[1] * Corresponding Author

users 24 hours a day at any location on the globe. This is the era where every academic library is transforming into at least a hybrid library and some of them even maintaining virtual existence, offering their services even beyond the working hours through online library portals. The academic libraries in the past two decades have begun to acquire or subscribe to very large collections of digital resources. Thanks to nationwide initiatives carried out by the Ministry of Human Resource Development under National Mission on Education through Information and Communication Technology, the University Grants Commission and Information Library Networks (INFLIBNET) center have come up with UGC Info Net, NLIST and Sakshat and many other e-portals through which several thousand e-books and e-journals are made available to the student community.

2. E-Resources and the Language Barrier:

Academic libraries have a very pivotal role to play not only in the academic development of the students but also in the research and extension activities. Their contribution is immense in the national building activity. To provide proper educational and research environment, the academic library must be able to offer world class resources, latest development in a subject area, new trends and current research outcomes. This is made possible thanks to the online availability of e-resources both as free and also at a premium. Thanks to the initiatives of INFLIBNET, a vast number of e-resources published by top academic publishers are made available to the students and teaching faculty at a substantially discounted price through library consortia. Not only that, academic libraries are able to access a very large number of e-journals on the same day of their publication. This is helping the researchers, teachers and the student community to keep abreast with the latest developments in their respective fields. Many portals such as Sakshat, CEC, NPTEL etc., are also developed by the union government, that feature world class learning resources, including video lectures by eminent academicians.

One of the most fundamental barriers that stops the learners of higher education from favoring the e-resources to conventional books is the language barrier. A great deal of world class literature in every field of knowledge is produced and made available in English, a language the rural populace in India are not comfortable with. India is a land of multiple languages. Though the national language of the country is Hindi, more than 60% the country's population has many other vernacular languages as its mother tongue. A great number of youth in the rural and semi-urban areas still pursue their education – even at higher education levels – in their mother tongues. Owing to this, their comprehension levels of English language are as rudimentary as their computer handling skills. This compels them to bank on books published by local publishers in the language of land, which in turn are way below the international standard in terms of quality and concurrence. This results in a vicious circle where the knowledge acquired is obsolete and so does not reflect the modern developments, as translation into vernacular language with ease, dexterity and efficiency reflecting the purport of the original text without losing the grace, is only seldom seen. This makes the rural youth a second rate citizens in knowledge acquisition, alienating them further. The facility of translation enjoyed by their European counterparts and even other Asians such as Chinese and Japanese is also not available to them. While most websites offer the option of viewing their content in multiple international languages and so sometimes offer them in the official national language of India, Hindi, they would not offer their content in other major Indian languages such as Tamil, Telugu, Marathi or Bengali. Since the learners want the accessing technologies as well as the learning material in their native languages, this leaves them out of bounds.

Another disheartening fact is that the societies in developing countries are mostly the end users, not the producers of new knowledge. And even the marginal contribution by the new knowledge producers from the developing world is always invariably in English as they want their inventions and discoveries to be heard by larger international audience.

Thus any modern addition to the existing body of knowledge is always in international languages and so the users of other languages are deprived from the latest contributions. Another major hurdle is keying in the search words in vernacular languages which, however, is considerably reduced by the introduction of Unicode writing system, but is still slightly technical and not fully intuitive. While the permutations and combinations of just 26 letters produce all the words and expressions in English, the phonetic languages of India have 56 letters with their hraswas and *dirghas* and *dwitwaksharas* thus making the keying-in a far more complex task.

3. Language Translation: International and Indian Scenarios

There has been some serious research undertaken by several researchers in tackling the issue of linguistic impediments and many government funded agencies made attempts to surmount the challenge. Different techniques and approaches have been evolved by various research groups for devising machine translation systems for translating text from one language to others. Natural Language Processing (NLP) and Machine Translation (MT) are widely studies in the field of computational linguistics under the broader field of artificial intelligence.

Logos Corporation has developed one of the most successful machine translation systems named 'The Logos', which provides reasonably appreciable automatic translation of texts, which can then be improved through human intervention. The system is widely popular as a first level draft generation mechanism among translation bureaus across the globe, owing to the scope it offers to add domain specific technical terminology and its adaptability to new linguistic constructs. The Logos was evaluated by Douglas W. Oard and Paul Hackett (1997) which revealed that in the instances where translation is to be rendered on a relatively large number of documents, Machine Assisted translation technologies such as the ones developed by Logos Corporation come very handy. They concluded that in some specific instances document translation may prove to be more efficacious than the query translation method.

J. Scott McCarley, (1999) also considered the problem of whether document translation is superior to query translation. He investigated the translation efficiency of both the methods between English and French languages and vice versa, and compared them with hybrid systems that incorporate both the approaches. He concluded that hybrids of document and query translation-based systems are far more superior to systems based on query translation. His study revealed that hybrid systems even significantly perform better than human-quality query translation systems.

The C-DAC Mumbai has been involved in a project titled 'English to Indian Languages: Machine Translation System (E-Ilmt)'. Also known as Anuvadaksh, this project aims to design and deploy a Machine Translation System from English to Indian Languages in Tourism, Agriculture and Healthcare domains. This project, started in the year 2006, is funded by the Ministry of Communication and Information Technology, Government of India and focuses on breaking the language barrier that plagues a multilingual country like India. Since majority of the Indian population is not familiar with English whereas most of the information available on web or electronic information is in English, the need for an automatic language translator to reach out to the common man is realized. C-DAC Mumbai’s primary objective was to develop statistical models and resources for a statistical Machine Translation system for translating documents from English into major languages of India such as Hindi, Marathi and Bengali languages. Of late, Oria, Tamil, Urdu, Gujarati and Bodo languages have also been added. Statistical model of Machine Translation is based on mathematical models of human translation process and renders the output in a way that is very similar to human translation. The tool takes horrendously long time to translate even a few lines into the selected Indian language. Even the quality of translation is very unsatisfactory. A sample translation by ANUVADAKSH is presented in the subsequent part of this paper. R. Ananthakrishnan and others (2008) maintain that simple syntactic and morphological processing can help improve English-Hindi statistical machine translation systems and suggest ways and means to achieve this.

For automatic machine based translation of one Indian language into other, the Indian Language Technology Proliferation and Deployment Centre has designed a tool named Sampark. The project is funded by Technology Development for Indian Languages Program of the Department of Information Technology, Government of India and is developed by a Consortium of Institutions ranging from IITs, IIITs, CDACs, IISc and other universities. The tool is still in development stage and produces very poor and faulty translations.

Efforts to develop machine aided Natural Language Processing (NLP), began way back in 1996 at IIT Bombay thanks to the financial assistance it received from the United Nations University, Tokyo. The focus of this project was to devise and develop Universal Networking Language, a multilingual information exchange system for the web. The project was collaborated by as many as fifteen research groups across the globe. Later in the year 2000, the Department of Information Technology set up Center for Indian Language Technology (CFILT) at IIT Bombay to augment its deep semantics and multiple language efforts. The emphasis on semantics resulted in some appreciable developments.

Nair Latha R and David Peter S. (2012) underscore the heavy demand for translation of documents from one language to another language in India. They point out that despite several concerted efforts to develop machine translation in the past couple of decades by various agencies, efforts have yielded very limited results in developing a fully automatic high quality machine translation system. In their paper they discussed various approaches which have been applied in translation systems for Indian languages. Some of the important Indian language translation systems implemented with these techniques along with their capabilities and limitations were also discussed.

Sanjay Kumar Dwivedi and Pramod Premdas Sukhadeve (2010) have described the efforts undertaken in India in this direction. Pointing out that the official correspondence of the central government of India is either in English or Hindi whereas the state governments prefer to work in their provincial languages, they underscore the high demand for translation of documents from one language to another language in a linguistically diverse country like India.

Salil Badodekar (2003) Indian Institute of Technology, Mumbai studied 'Translation Resources, Services and Tools for Indian Languages' and enlisted various translation services and tools available in India.

N. Swapna,, N.Hareenkumar, and B. Padmaja Rani (2012) point out that people invariably use every day one or other form of modern information retrieval system such as Google or specially library-customized systems. In traditional Information Retrieval method, the query and the retrieved documents are in the same language. The traditional Information Retrieval systems consider foreign language documents as unwanted 'noise'. Hence there is a need to develop Information Retrieval systems that retrieve all related documents in every language irrespective of the queried language. Hence the focus should be on developing the bilingual, cross-lingual and multi-lingual Information Retrieval systems. In India where many languages are spoken by large number of people, the user prefers to query in one language and expects to retrieve suitable documents in another language. All the Indian languages are phonetic based and are represented by syllable as a basic linguistic unit. Translation can be done at query level or at the retrieved document level. Dictionaries, machine translation systems and parallel corpora are basically relied on while translating one language into another. While query translation typically, relies on either dictionary based or corpus based translation, machine translation is mostly used for document translation.

Bandyopadhyay, Sivaji&Naskar, Sudip (2005) point out to a survey that reveals that the Machine Translation software developed in India for translation from English to Indian languages and among Indian languages are used in field testing or are available as web translation service. They inform that these systems developed by the Indian agencies are also used for imparting knowledge on machine translation techniques and tools to the students and researchers in various institutions. The most common translation areas are government documents and news reports.

The development of machine translation systems post Second World War is described by Antony P. J (2013) who also summarizes various machine translation efforts carried out by Indian agencies. He also contends that developing a full-fledged natural language Machine Translation system is tremendously a challenging task. He deliberates on two important approaches adopted by the researchers in this regard - rule-based as well as statistical-based approaches. He points out that while only about 5% of the world's population speaks English as a first language it is nevertheless widely used in media, trade, science and technology, and education. Thus arguably, there is great demand for translation systems that effectively translate documents in English language into other languages. He also concurs that though Machine Translation in India was initiated more than two decades ago, it is still at a very basic stage.

4. The Indian Effort to Surmount the Language Barrier

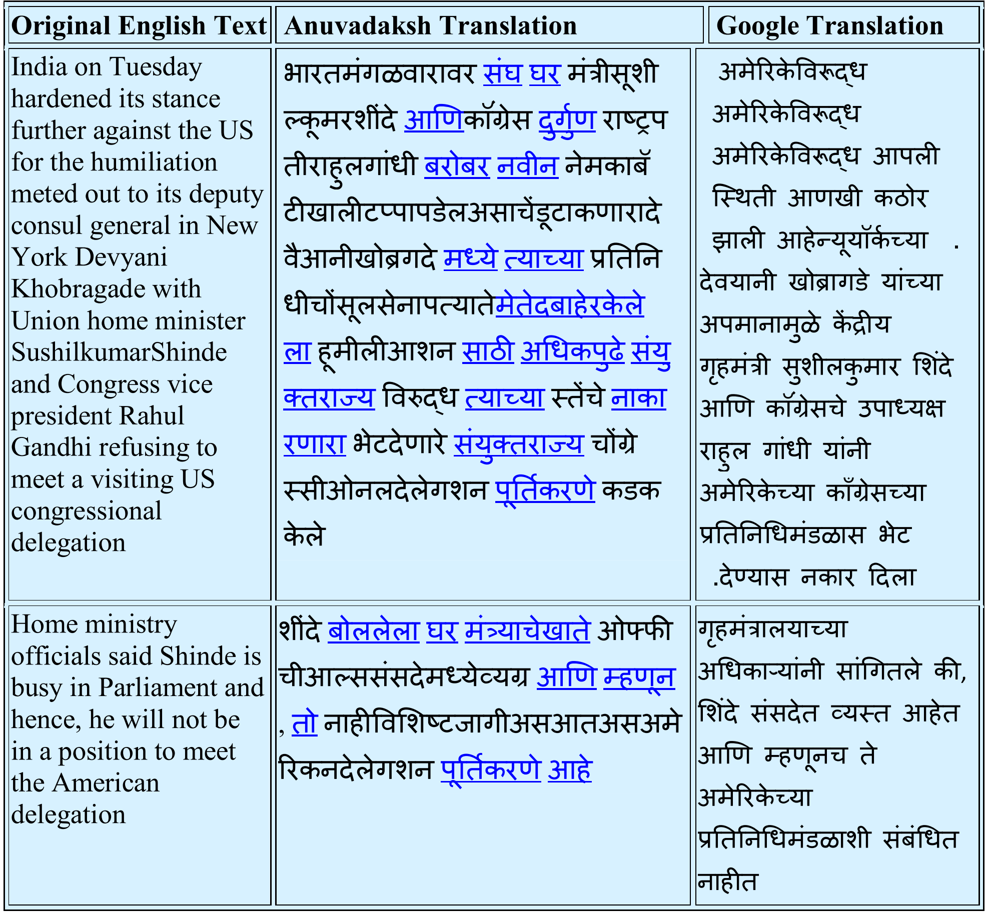
There have been some efforts by the premier technology universities of India like Indian Institute of Technology, Kanpur, CDAC, Pune and Bangalore who produced translation software such as Anglabharati, MaTra and Mantra. A brief list of Machine Translation systems developed by various Indian agencies is herewith provided:

Table 1: A Brief Overview of Some Prominent Indian Machine Translation Systems.

Systems	Year	Organization/ Institute	Description
AnglaBharti	1991	IIT Kanpur	It is a general purpose system with a Lexicon that has customizable domain.
Anusaaraka	1995	IIT Kanpur	This system is mainly used to translate stories for children between Indian languages. It aspires for perfect preservation of information.
Mantra	1999	C-DAC, Bangalore	The system was primarily developed for the Rajya Sabha Secretariat. Works on Indian language pairs.
MaTra	2004	C-DAC, Mumbai	It has a text categorization component which breaks up the complex English sentences into simpler sentences which then analyzes and produces Hindi translation of the original English text.
AnuBharti	2004	IIT Kanpur	Translation from English to Indian Languages and translation within Indian languages.
Shiva and Shakti	2004	Carnegie Mellon University USA, IIIT Hyd and Institute of Science, Bangalore	The Shiva is a machine translation system for translating English sentences into other languages. Shakti system on the other hand is designed to produce machine translation systems for Hindi, Marathi and Telugu.
Anubaad	2004	Jadavpur University, Kolkata	Anubaad a hybrid MT system for translating English news headlines into Bengali. At present, the system is capable of working at the sentence level.
Sampark	2009	Consortium of	Machine translation system among Indian

		Institutions	languages developed by the Consortium of institutions. Currently experimental systems have been released namely {Punjabi, Urdu, Tamil, Marathi} to Hindi and Tamil-Hindi Machine Translation systems.

But despite spending millions of rupees in the development of translation software with an intention to unite the linguistically diverse country with one thread, the show has been a poorly performed one. These translation software fail to recognize the semantic intricacies of language and deliver the output that sometimes verges on ridiculousness. While Indian agencies still struggle to provide a decent translation platform to the Indian users, the search engine giant Google has developed its translation tool for major international languages including eleven Indian languages. Here is a comparative analysis of efficiency of translation by Anuvadaksh and Google Translate services:



Original English Text	Anuvadaksh Translation	Google Translation
India on Tuesday hardened its stance further against the US for the humiliation meted out to its deputy consul general in New York Devyani Khobragade with Union home minister SushilkumarShinde and Congress vice president Rahul Gandhi refusing to meet a visiting US congressional delegation	भारतमंगळवारावर संघ घर मंत्रीसूशील्कूमरशींदे आणिकॉंग्रेस दुर्गुण राष्ट्रपतीराहुलगांधी बरोबर नवीन नेमकाबॅटीखालीटप्पापडेलअसाचेंडूटाकणारादेवैआनीखोब्रगदे मध्ये त्याच्या प्रतिनिधीचोंसूलसेनापत्यातेमेतेदबाहेरकेलेला हूमीलीआशन साठी अधिकपुढे संयुक्तराज्य विरुद्ध त्याच्या स्तेंचे नाकारणारा भेटदेणारे संयुक्तराज्य चोंग्रेस्सीओनलदेलेगशन पूर्तिकरणे कडक केले	अमेरिकेविरूद्ध अमेरिकेविरूद्ध अमेरिकेविरूद्ध आपली स्थिती आणखी कठोर झाली आहेन्यूयॉर्कच्या . देवयानी खोब्रागडे यांच्या अपमानामुळे केंद्रीय गृहमंत्री सुशीलकुमार शिंदे आणि कॉंग्रेसचे उपाध्यक्ष राहुल गांधी यांनी अमेरिकेच्या कॉँग्रेसच्या प्रतिनिधिमंडळास भेट .देण्यास नकार दिला
Home ministry officials said Shinde is busy in Parliament and hence, he will not be in a position to meet the American delegation	शींदे बोललेला घर मंत्र्याचेखाते ओफ्फीचीआल्ससंसदेमध्येव्यग्र आणि म्हणून , तो नाहीविशिष्टजागीअसआतअसअमेरिकनदेलेगशन पूर्तिकरणे आहे	गृहमंत्रालयाच्या अधिकाऱ्यांनी सांगितले की, शिंदे संसदेत व्यस्त आहेत आणि म्हणूनच ते अमेरिकेच्या प्रतिनिधिमंडळाशी संबंधित नाहीत

Figure 1: English to Marathi translation using ANUVADAKSH developed by Technology Development for Indian Languages, A Government of India Enterprise and Google Translate service. The image shows the horrendous quality of translation.

5. Conclusion

Instead of spending millions of Rupees on developing only infrastructure, what is urgently needed is the development and deployment of automatic parsers and translation software as a plug in for web browsers. All the library websites and portals of institutions of higher education shall be mandated to incorporate translation plug ins on their home page. The Nlist, INDEST, NPTEL and other portals developed by the government of

India shall be asked to provide translation to their e-resources in Indian languages. This will enable the users from rural and semi-urban regions to familiarize themselves with what is available for them out there in the electronic environment.

6. References & Bibliography:

6.1 Journal Article

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

[2] Nair, Latha R and David Peter S., Machine Translation Systems for Indian Languages, the *International Journal of Computer Applications* 39(1), **(2012)**, pp. 24-31.

[3] N .Swapna, et al., Information Retrieval In Indian Languages: A Case StudyOn Cross-Lingual And Multi-Lingual, *International Journal of Research in Computer and Communication technology*, IJRCCT, ISSN 2278-5841, Vol 1, No. 4, **(2012)**.

[4] P. J Antony, Machine Translation Approaches and Survey for Indian Languages**,** *Computational Linguistics and Chinese Language Processing* Vol. 18, No. 1, The Association for Computational Linguistics and Chinese Language Processing**, (2013),** pp. 47-78.

[5] Quadri, Ganiyu Oluwaseyi, Impact of ICT Skills on the Use of E-Resources by Information Professionals: A Review of Related Literature, *Library Philosophy and Practice*, June, **(2012)**.

[6] Badodekar, Salil. Translation Resources, Services and Tools for Indian Languages, Computer Science and Engineering Department Indian Institute of Technology, Mumbai,**(2003).**

6.3 Conference Proceedings

[7] Bandyopadhyay, Sivaji, Naskar, Sudip,Use of Machine Translation in India: Current Status, *Proceedings of MT SUMMIT X*, Phuket, **(2005),** pp. 465-470.

[8] Oard Douglas W. and Hackett Paul, Document Translation for Cross_Language *The Sixth Text REtrieval Conference (TREC-6)*. Text Retrieval at the University of Maryland, U.S. Dept. of Commerce, Technology Administration, National Institute of Standards and Technology, Information Technology, **(1997).**

[9] McCarley J. Scott, Should we Translate the Documents or the Queries in Cross-language Information Retrieval?, *ACL '99: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, Association for Computational Linguistics, **(1999).**

[10] R. Ananthakrishnan, Jayprasad Hegde, Pushpak Bhattacharyya, Ritesh Shah and M. Sasikumar, Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation, International Joint Conference on NLP (IJCNLP08), Hyderabad, India, **(2008),** January 10-12.

6.4 Web Resources

[11] http://www.cdacmumbai.in/index.php/cdacmumbai/research_publications/projects/development_of_english_to_indian_languages_m

[12] http://www.nlist.inflibnet.ac.in

[13] http://www.nptel.ac.in

[14] http://paniit.iitd.ac.in/indest/

[15] http://sakshat.ac.in