Use of Machine Translation in India: Current Status

Sudip Naskar
Computer Science and Engineering Department
Jadavpur University
Kolkata, India, 700032
sudip_naskar@hotmail.com

Sivaji Bandyopadhyay
Computer Science and Engineering Department
Jadavpur University
Kolkata, India, 700032
sivaji_cse_ju@yahoo.com

Abstract

A survey of the machine translation systems that have been developed in India for translation from English to Indian languages and among Indian languages reveals that the MT softwares are used in field testing or are available as web translation service. These systems are also used for teaching machine translation to the students and researchers. Most of these systems are in the English-Hindi or Indian language-Indian language domain. The translation domains are mostly government documents/reports and news stories. There are a number of other MT systems that are at their various phases of development and have been demonstrated at various forums. Many of these systems cover other Indian languages beside Hindi.

1 Introduction

India has a diverse list of spoken languages. At least 30 different languages and around 2000 dialiects have been identified. The Constitution of India has stipulated the usage of Hindi and English to be the two languages of official communication for the national government. Additionally, it classifies a set of 22 *scheduled languages* which are languages that can be officially adopted by different states for administrative purposes, and also as a medium of communication between the national and the state governments, as also for examinations conducted for national government service. The 22 *scheduled languages* are Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu. English is very widely used in the media, commerce, science and technology and education. Many of the states have their own regional language, which is either Hindi or one of the other scheduled languages. Only about 5% of the population speak English.

In a large multi-lingual society like India, there is a great demand for translation of documents from one language to another. Most of the state governments work in the respective regional languages whereas the Union Government's official documents and reports are in bilingual form (Hindi/English). In order to have a proper communication there is a need to translate these documents and reports in the respective regional languages. The newspapers in regional languages are required to translate news in English received from International News Agencies. With the limitations of human translators most of this information (reports and documents) is missing and not percolating down. A machine assisted translation system or a translator's workstation would increase the efficiency of the human translators.

As is clear from above, the market is largest for translation from English into Indian languages, primarily Hindi. Hence, it is no surprise that a majority of the Indian Machine Translation (MT) systems have been developed for English-Hindi translation.

Machine Translation activities in India are relatively young. The earliest efforts date from the mid 80s and early 90s. The prominent among these efforts are the research and development projects at Indian Institute of Technology (IIT) Kanpur (http://www.cse.iitk.ac.in/users/rmk/proj/proj.html#mt), Computer and Information Sciences Department, University of Hyderabad (http://www.uohyd.ernet.in/), National Center for Software Technology (NCST), Mumbai (now, Center for Development of Advanced Computing (CDAC), Mumbai) (http://www.ncst.ernet.in/) and Center for Development of Advanced Computing (CDAC), Pune (http://www.cdac.in/). The Department of Information Technology, Ministry of Communications and Information Technology, Government of India, through its Technology Development in Indian Languages (TDIL) Project (http://tdil.mit.gov.in) since 1990-'91 and Department of Official Languages, Ministry of Home Affairs, Government of India, since 1989-90 have played instrumental roles by funding these projects. The following domains have been identified for the development of domain specific translation systems: government administrative procedures and formats, parliamentary questions and answers, pharmaceutical information, legal terminology and important judgments, and so on.

Since the mid and late 90's, a few more projects have been initiated—at Indian Institute of Technology (IIT) Bombay (http://www.iitb.ac.in/), International Institute of Information Technology (IIIT) Hyderabad (http://www.iiit.net/), Anna University – KB Chandrasekhar Research Center (AU-KBC) Chennai (http://www.au-kbc.org/) and at the Computer Science and Engineering Department, Jadavpur University, Kolkata (http://www.jadavpur.edu/).

There are also a couple of efforts from the private sector - from Super Infosoft Private Limited, and more recently, the IBM India Research Laboratory.

A survey of the machine translation systems that have been developed in India for translation from English to Indian languages and among Indian languages reveals that the MT softwares are used in field testing or are available as web translation service. These systems are also used for teaching machine translation to the students and researchers. A report on the use of these machine translation systems for teaching in Indian Universities has been presented in (Bandyopadhyay, 2002). Most of these systems are in the English-Hindi or Indian language-Indian language domain with exceptions of a Hindi-English and an English-Kannada machine translation system. The translation domains are mostly government documents/reports and news stories. There are a number of other MT systems that are at their various phases of development and have been demonstrated at various forums. Many of these systems cover other Indian languages beside Hindi.

2 MT Systems in Field Testing or as Web Service

The following machine translation systems are either being used in field-testing or are available as a web service. Most of these systems are in the English-Hindi language domain with exceptions of a Hindi-English and an English-Kannada machine translation system.

2.1 ANGLABHARTI and ANUBHARTI Technology

The ANGLABHARTI project was launched by Professor R. M. K. Sinha at the Indian Institute of Technology, Kanpur in 1991 for machine aided translation from English to Indian languages. Professor Sinha has pioneered MT research in India. The system's approach and lexicon is general-purpose with provision for domain customization. The system has been applied in several domains such as public health compaign, routine office-correspondance, technical-mannual etc.

The first prototype was built for English to Tamil in 1991 and later a more comprehensive system was built for English to Hindi translation. AnglaBharti is a pattern directed rule based system with context free grammar like structure for English (source language). It generates a 'pseudo-target' (Pseudo-Interlingua) applicable to a group of Indian languages (target languages) such as Indo-Aryan family (Hindi, Bengali, Assamese, Punjabi, Marathi, Oriya, Gujarati etc.), Dravidian family (Tamil, Telugu, Kannada & Malayalam) and others. A set of rules obtained through corpus analysis is used to identify plausible constituents with respect to which movement rules for the 'pseudo-target' is constructed. Within each group the languages exhibit a high degree of structural homogeneity. The similarity has been exploited to a great extent in the system. A language specific text-generator converts the 'pseudo-target' code into target language text. Paninian framework based on Sanskrit grammar using Karak (similar to case) relationship provides an uniform way of designing the Indian language text generators. They also use an example-base to identify noun and verb phrasals and resolve their semantics. An attempt is made to resolve most of the ambiguities using ontology, syntactic & semantic tags and some pragmatic rules. The unresolved ambiguities are left for human post-editing. Some of the major design considerations in design of AnglaBharti have been aimed at providing an uniform mechanism by which translation from English to majority of Indian languages with attachment of appropriate text generator modules becomes feasible. The translation system has also been interfaced with text-to-speech module and OCR input.

The English to Hindi version named *AnglaHindi*, based on the AnglaBharti machine aided translation system has been web-enabled and is available at (http://anglahindi.iitk.ac.in). The technical know-how of this technology has been transferred on a non-exclusive basis to Center for Deveopment of Advanced Computing (CDAC), Noida for commercialization.

The AnglaBharti technology has also been transferred to eight different organizations under AnglaBharti Mission for development of Machine Aided Translation (MAT) systems for English to different Indian languages catering to 12 regional languages of the country. Under this mission, IIT Mumbai will be working on Marathi & Konkani and will be developing AnglaMarathi & AnglaKonkani; IIT Guwahati will be working on Asamiya (Assamese) & Manipuri and will be developing AnglaAsamiya & AnglaManipuri; CDAC Kolkata will be working on Bangla (Bengali) and will be developing AnglaBangla; CDAC(GIST group) Pune will be working on Urdu, Sindhi & Kashmiri and will develop AnglaUrdu, AnglaSindhi & AnglaKashmiri; CDAC Thiruvananthpuram will be working on Malyalam and will be developing AnglaMalayalam; Thapar Institute of Engineering and Technology (TIET) Patiala will be working on Punjabi and will be developing AnglaPunjabi; Jawaharlal Nehru University (JNU) New Delhi will be working on Sanskrit and will be developing AnglaSanskrit and Utkal University Bhuvaneshwar will be working on Oriya and will be developing AnglaOriya.

In 1995, Prof. Sinha developed the ANUBHARTI methodology that follows the example based machine translation strategy. This methodology has been used for Hindi to English translation. The AnuBharti approach works more efficiently for similar languages such as among Indian languages. In such cases the word-order remains the same and one need not have pointers to establish correspondences.

Both of these system architectures, AnglaBharti and AnuBharti, have undergone a considerable change from their initial conceptualization. In 2004, phase-II of the system development has been launched which addresses many of the shortcomings of the earlier architectures. These are named AnglaBharti-II and AnuBharti-II. Both these systems are hybridized with varying degree of hybridization of different paradigms.

AnglaBharti-II uses a generalized example-base (GEB) for hybridization besides a raw example-base (REB). During the development phase, when it is found that the modification in the rule-base is difficult and may result in unpredictable results, the example-base is grown interactively by augmenting it. At the time of actual usage, the system first attempts a match in REB and GEB before invoking the rule-base. In AnglaBharti-II, provision have been made for automated pre-editing & paraphrasing, generalized & conditional multi-word expressions, recognition of named-entities, domain customization tools and incorporated an error-analysis module and statistical language-model for automated post-editing.

The AnuBharti strategy has now been generalized in AnuBharti-II to cater to Hindi as source language for translation to any other language, though the generalization of the example-base is dependent upon the target language. The core of AnuBharti-II architecture is a generalized hierarchical example-base. Development of the example-base for Hindi to other Indian languages is lot easier as compared to a dissimilar language. Hindi like all other Indian languages is a relatively free word-group order language. The input Hindi sentence is converted into a standardized form by shallow grammatical analysis to take care of word-order variations. The standardized Hindi sentence is matched with a top level standardized example-base. In case no match is found then a shallow chunker is used to fragment the input sentence into units that are then matched with a hierarchical example-base. The translated chunks are positioned by matching with sentence level example base. An error-analysis module and statistical language model on lines similar to AnglaBharti-II are also being incorporated. Human post-editing is performed primarily to introduce determiners that are either not present or difficult to estimate in Hindi.

Besides these, they are also currently engaged in development of translation system for bi-lingual text in Hinglish (Hindi mixed with English) and system for speech to speech translation.

The contact person is Prof. R M K Sinha (rmk@iitk.ac.in).

2.2 MANTRA MT System

MANTRA (MAchiNe assisted TRAnslation tool) translates English text into Hindi in a specified domain of personal administration, specifically gazette notifications, office orders, office memorandums and circulars. *Mantra* uses *Lexicalized Tree Adjoining Grammar (LTAG)* to represent the English as well as the Hindi grammar. It uses *Tree Adjoining Grammar (TAG)* for parsing and generation. The input to the *Mantra* system can be through text documents, outputs of speech recognition programs or OCR packages. The *Mantra* has become part of “The 1999 Innovation Collection” on Information Technology at Smithsonian Institution’s National Museum of American History, Washington DC, USA. Further details about the *Mantra* system can be obtained from the C-DAC website (http://www.cdac.in/html/aai/mantra.asp).

Initially, the *Mantra* system was started with the translation of administrative document such as appointment letters, notification, circular issued in Central government from English to Hindi. Currently, the system is being expanded to cover domains of finance, agriculture, healthcare, information technology, education and finally to the general purpose activities in the Governmental Domain. This project namely *MANTRA-Rajbhasha* based on the *MANTRA* technology developed by C-DAC has been funded by the Department of Official Language, Ministry of Home Affairs, Government of India.

The grammar is specially designed to accept, analyze and generate sentential constructions in "officialese" domain. Similarly, the lexicon is suitably restricted to deal with meanings of English words as used in its subject-domains. The system is ready for use in its domains.

The system namely *MANTRA-Rajyasabha* is developed for the Rajya Sabha Secretariat, the Upper House of Parliament of India. It translates 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. This project was funded by the Rajya Sabha Secretariat.

Recently, work has been initiated on other language pairs such as English-Bengali, English-Telugu, English-Gujarati, Hindi-English and among India languages such as Hindi-Bengali, and Hindi-Marathi.

The contact person is Dr Hemant Darbari (darbari@cdac.in).

2.3 ANUSAARAKA MAT System

A machine aided translation system (ANUSAARAKA) among Indian languages has been built with funding from the TDIL project. *Anusaaraka* is a Language accessor rather than a machine translation system in true sense. *Anusaarakas* have been built from Telugu, Kannada, Bengali, Marathi, and Punjabi to Hindi. Alpha versions of all of these have been released so that their field testing can be carried out. The beta-versions are expected to be released soon.

It is domain free but the system has mainly been applied for translating children's stories. An e-mail server been established for the *Anusaarakas*. It currently holds alpha-versions of *Anusaarakas* for translating Telugu, Kannada, Marathi, Bengali, and Punjabi into Hindi. To run the *Anusaaraka* on a given text, e-mail has to be sent to (nandi@anu.uohyd.ernet.in) with the name of the language in the subject line. For example, if 'Telugu' is put in the subject line, it automatically runs the Telugu to Hindi *Anusaaraka*. The output produced is sent back to the sender. The computer keeps a copy of the output for a later study of results. The text should be in 7-bit ISCII coding. Similarly help by mail is available if mail is sent with subject 'Help'. The systems with complete source code and language data are available as "free" software and can be downloaded from the website of the Language Technologies Research Center, Indian Institute of Information Technology, Hyderabad (http://ltrc.iiit.net/). Further details about the *Anusaaraka* system can be obtained from their website (http://www.iiit.net/ltrc/Anusaaraka/anu_home.html).

The focus in Anusaaraka is not mainly on machine translation, but on Language Access between Indian languages. Using principles of Paninian Grammar (PG), and exploiting the close similarity of Indian languages, an Anusaaraka essentially maps local word groups between the source and target languages. Where there are differences between the languages, the system introduces extra notation to preserve the information of the source language. Thus, the user needs some training to understand the output of the system. The project originated at IIT Kanpur, and later shifted mainly to the Centre for Applied Linguistics and Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad. It was funded by TDIL. It is now being carried out at the Language Technologies Research Center, Indian Institute of Information Technology, Hyderabad with financial support from Satyam Computers Private Limited. They are currently attempting an English-Hindi Anusaaraka MT system

The contact person is Prof. Rajeev Sangal (sangal@iiit.net).

2.4 SHIVA and SHAKTI MT System

Two machine translation systems from English to Hindi, *Shiva* and *Shakti* are being developed jointly by Carnegie Mellon University USA, Indian Institute of Science, Bangalore, India, and International Institute of Information Technology, Hyderabad. The two experimental systems have been released for experiments, trials, and user feedback. The Example-based Machine Translation system (*Shiva*) can be run from the site (http://ebmt.serc.iisc.ernet.in/mt/login.html) and the machine translation system (*Shakti)* can be run from its own site (http://shakti.iiit.net). Currently Shakti(Version 0.81, New Release - 18th April, 2005) is working for three target languages (Hindi, Marathi and Telegu). The user has to log on to the system for translating English sentences into the appropriate language. The system asks the user to choose a font from a list and shows the translated output in the chosen font. If the user is not satisfied with the system generated translation, the system asks for meanings of various source language components (word, phrase and even sentence level) from the user and retranslates the input text.

2.5 UNL-based English-Hindi MT System

The Indian Institute of Technology, Bombay has developed a machine translation system for translating from English to Hindi using the Universal Networking Language (UNL) as the interlingua. The UNL is an international project of the United Nations University, with an aim to create an Interlingua for all major human languages. IIT Bombay is the Indian participant in UNL. The MT system can be used at the site (http://www.cfilt.iitb.ac.in/machine-translation/eng-hindi-mt). Two other demo systems can also be run from the site: one for Hindi-UNL and another for UNL-Hindi conversion. They are also working on MT systems from English to Marathi and Bengali using the UNL formalism. The contact person is Prof. Pushpak Bhattacharyya (pb@cse.iitb.ac.in).

2.6 MATRA MT System

The Natural Language group of the Knowledge Based Computer Systems (KBCS) division at the National Centre for Software Technology (NCST), Mumbai (currently CDAC, Mumbai) is working on *MaTra*, human-aided transfer-based translation system for English to Hindi. The work is supported under the TDIL Project. The *MaTra* lexicon and approach is general-purpose, but the system has been applied mainly in the domains of news, annual reports and technical phrases, and has been funded by TDIL.

A Hybrid approach system *Vaakya* has been developed at NCST, Bombay for translating English news stories to Hindi. The system can handle single verb sentences. Prototype *Vaakya* system is now being enhanced and adapted for providing web translation service to the news agencies from English news stories to Hindi. More details about the *MaTra* system can be obtained from the website of the NCST (http://www.ncst.ernet.in/matra/) They are using the translation system in a project on Cross Lingual Information Retrieval (CLIR) that enables a person to query the web for documents related to health issues in Hindi. Details about the CLIR system can be seen in (http://www.ncst.ernet.in/projects/clir/). The contact person is Ms. Kavitha Mohanraj (kavitham@ncst.ernet.in).

2.7 English-Kannada MT System

The Computer and Information Sciences Department at the University of Hyderabad has worked on an English-Kannada MT system, using the Universal Clause Structure Grammar (UCSG) formalism. This is essentially a transfer-based approach, and has been applied to the domain of government circulars, and funded by the Karnataka government. Further details about the English-Kannada machine aided translation system can be obtained from their website (http://www.languagetechnologies.ac.in/lerc/mat/mat.htm). Currently, efforts are on to improve the system and apply it for English-Telugu translation as well. The contact person is Prof. K Narayana Murthy (knmcs@uohyd.ernet.in).

2.8 Tamil-Hindi MAT System

The Anna University KB Chandrasekhar Research Centre (http://www.au-kbc.org/) at Chennai has developed a Tamil-Hindi machine aided translation system which can be accessed at (http://www.au-kbc.org/research_areas/nlp/demo/mat/). There is a keyboard for input of the Tamil Text and the output can be seen in a Hindi font downloadable from the site. The translation system is *Anusaaraka* based and work on an English-Tamil human aided machine translation system is going on. The contact person is Prof. CN Krishnan (cnkrish@au-kbc.org).

2.9 ANUVAADAK MT System

Super Infosoft Private Limited, Delhi have developed a general purpose English-Hindi machine translation tool *Anuvaadak 5.0* that supports post-editing. It has inbuilt *dictionaries for specific domains* e.g. official, formal, agriculture, linguistics, technical, and administrative. When Hindi meaning of the English word is not available in dictionary, facility of transliteration is provided. The software runs on any operating system in the Windows family. A demonstration version of the system can be downloaded from the website (http://www.mysmartschool.com/pls/portal/portal.MSSStatic.ProductAnuvaadak). The contact person is Mrs. Anjali Rowchowdhury (anjalir@ndf.vsnl.net.in).

3 Other MT Systems

The Computer Science and Engineering Department of Jadavpur University (http://www.jadavpur.edu/) is working on an machine translation system *ANUBAAD* for translating English news items to Bengali using a phrasal Example Based Machine Translation (EBMT) approach. The current version of the system works at the sentence level. A separate activity is going for developing a semantics-based EBMT system for translating English news headlines to Bengali. A demonstration version of the system has been developed. The English-Bengali MT system architecture has been used to develop a English – Hindi MT system that works at the simple sentence level. These machine translation systems are used for teaching the students and researchers who work in the area of machine translation. Recently, works on Bengali – English and Manipuri – English machine translation systems have been started. The contact person is Prof. Sivaji Bandyopadhyay (sivaji_cse_ju@yahoo.com).

Utkal University, Bhuvaneshwar is working on a English-Oriya machine translation system OMTrans (http://www.ilts-utkal.org/omtrans.htm). The contact person is Prof. Sanghamitra Mohanty (sangham1@rediffmail.com). The Department of Mathematics, Indian Institute of Technology, Delhi is working on an English-Hindi example based machine translation system. They have developed algorithms for identification of divergence for English to Hindi EBMT system and a systematic scheme for retrieval from the English-Hindi example base. The contact person is Professor Niladri Chatterjee (niladri@maths.iitd.ernet.in). The IBM India Research Lab at New Delhi has recently initiated work on statistical machine translation between English and Indian languages, building on IBM's existing work on statistical machine translation.

4 Conclusion

MT is relatively new in India – about two decades of R & D efforts are now showing the results. In comparison with MT efforts in Europe and Japan, which are at least 3 decades old, it would seem that Indian MT has a long way to go. However, this can also be an advantage, because Indian researchers can learn from the experience of their global counterparts.

Development of Speech-Speech machine translation system from English to Indian languages and among Indian languages is the goal of the TDIL project and the various resource centers under the TDIL project are working towards that. Works on developing machine translation systems for other languages are also being initiated. There are governmental as well as voluntary efforts under way to develop common lexical resources and tools for Indian languages like POS tagger, Semantically rich lexicons and Wordnets. The NLP Association of India, regular international conferences like International National Conference on Natural Language Processing (ICON) and Lexical Resource Egroups like (lr_egroup@iiit.ac.in) are consolidating and coordinating NLP and MT efforts in India.

The time has come when translation activities in Indian languages should look beyond English as the source or the target language. The countries in the Asia-pacific region are renewing their relations in the area of tourism and trade. Development of controlled language machine translation systems in these domains from Indian languages to the languages of the Asia-pacific region like Japanese, Chinese, Korean, Thai and others and vice-versa will foster the development of people-to-people contact in this region.

5 Acknowledgements

Our thanks go to Council of Scientific and Industrial Research, Human Resource Development Group, New Delhi, India for supporting Sudip Naskar under their Senior Research Fellowship Award (9/96(402)2003-EMR-I).

References

R.M.K. Sinha. 2005. *Integrating CAT and MT in AnglaBharti-II Architecture.* In "Proceedings of EAMT 2005", Budapest, Hungary.

Sivaji Bandyopadhyay. 2004. Use of Machine Translation in India. *AAMT Journal,* 36: 25-31.

R.M.K. Sinha and A. Jain. 2003. *AnglaHindi: An English to Hindi Machine-Aided Translation System*. In "Proceedings of MT SUMMIT IX", New Orleans, Louisiana, USA.

D. Gupta and N. Chatterjee. 2003. *Identification of Divergence for English to Hindi EBMT.* In "Proceedings of MT SUMMIT IX", New Orleans, Louisiana, USA.

Sivaji Bandyopadhyay. 2002. *Teaching MT – An Indian Perspective.* In "Proceedings of the 6th EAMT Workshop on Teaching Machine Translation", Manchester, UK, 13-22.

Shachi Dave, Jignashu Parikh and Pushpak Bhattacharyya. 2001. Interlingua-based English-Hindi Machine Translation and Language Divergence. *Journal of Machine Translation,* 16 (4): 251-304.

Durgesh Rao. 2001. *Machine Translation in India: A Brief Survey.* In "Proceedings of SCALLA 2001 Conference", Banglaore, India.

Sivaji Bandyopadhyay. 2000. State and Role of Machine Translation in India. *Machine Translation Review*, 11: 25-27.