Journal of Computer Science 6 (10): 1111-1116, 2010
ISSN 1549-3636
© 2010 Science Publications

Machine Translation System in Indian Perspectives

Sanjay Kumar Dwivedi and Pramod Premdas Sukhadeve
Department of Computer Science, Babasaheb Bhimrao Ambedkar University,
Lucknow, India

Abstract: Problem statement: In a large multilingual society like India, there is a great demand for translation of documents from one language to another language. **Approach:** Most of the state government works in there provincial languages, whereas the central government's official documents and reports are in English and Hindi. **Results:** In order to have an appropriate communication there is a need to translate these documents and reports in the respective provincial languages. Natural Language Processing (NLP) and Machine Translation (MT) tools are upcoming areas of study the field of computational linguistics. Machine translation is the application of computers to the translation of texts from one natural language into another natural language. It is an important sub-discipline of the wider field of artificial intelligence. **Conclusion/Recommendations:** There are certain machine translation systems that have been developed in India for translation from English to Indian languages by using different approaches. It is this perspective with which we shall broach this study, launching our theme with a brief on the machine translation systems scenario in India through data and previous research on machine translation.

Key words: Machine translation, computational linguistics, language processing

INTRODUCTION

As India is a large multilingual country, different states have different regional languages; hence for proper communication there is a need of machine translation. But in India the earliest efforts starts from the mid 80s and early 90s. In India several Institutes work on Machine Translation. The prominent Institutes are as follows:

- The research and development projects at Indian Institute of Technology (IIT), Kanpur
- National Centre for Software Technology (NCST) Mumbai (now, Centre for Development of Advanced Computing (CDAC), Mumbai
- Computer and information Sciences Department, University of Hyderabad
- Centre for Development of Advanced Computing (CDAC), Pune
- Ministry of Communications and Information Technology
- Government of India, through its Technology Development in Indian Languages (TDIL) Project

Above Institutes co-operate imperative role in the field of machine translation from the years ago. Most of the machine translation systems have been developed by these Institutes by using various domains. Many of the domains have been identified for the development of domain specific translation systems; parliamentary questions and answers, pharmaceutical information, government documents and notice. Various machine translation systems have been developed in India using various systems for language translation from English to Indian languages.

Machine translation systems for Indian languages: In India Machine Translation systems have been developed for translation from English to Indian Languages and from regional languages to regional languages. These systems are also used for teaching machine translation to the students and researchers. Most of these systems are in the English to Hindi domain with exceptions of a Hindi to English (Sinha and Thakur, 2005) and English to Kannada (Kumar and Murthy, 2006) machine translation system. English is a SVO language while Indian regional languages are SOV and are relatively of free word-order. The translation domains are mostly government documents, health, tourism, news reports and stories. A survey of the machine translation systems that have been developed in media for translation from English to

Corresponding Author: Sanjay Kumar Dwivedi, Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India

Indian languages and among Indian languages reveals that the machine translation software is used in field testing or is available as web translation service. Indian Machine Translation system (Naskar and Bandyopadhyay, 2002) are presented below; these systems are used to translate English to Hindi language.

Anusaaraka systems among Indian languages: As Anusaaraka (1995) project which started at IIT Kanpur, by Prof. Rajeev Sangal. And is now being continued at IIIT Hyderabad, was started with the explicit aim of translation from one Indian language to another. It was funded by Technology Development in Indian Languages (TDIL) and financial support from Satyam Computers Private Limited.

Anusaaraka's have been built from Telugu, Kannada, Bengali, Punjabi and Marathi to Hindi. It is domain free but the system has been applied mainly for translating children's stories. Anusaaraka aims for perfect "information preservation". In fact, Asnusaaraka output follows the grammar of the source language (where the grammar rules differ and cannot be applied with 100% confidence).

For Example, a Bengali to Hindi Anusaaraka can take a Bengali text and produce output in Hindi which can be understood by the user but will not be grammatically perfect.

For example, for 80% of the Kannada words in the Anusaaraka dictionary (Bharati *et al*., 1997) of 30,000 root words, there is a single equivalent Hindi word which covers the senses of the original Kannada word. An e-mail server been established for the Anusaaraka's. To run the Anusaaraka on a given text, e-mail has to be sent with the name of the language in the subject line. For example, if 'Telugu' is put in the subject line, it involuntarily runs the Telugu to Hindi Anusaaraka. The focus in Anusaaraka is not mainly on machine translation, but on language access between Indian languages. Anusaaraka systems can be obtained from their website (http://www.iiit.net/ltrc/Anusaaraka/anu_home.html) they are currently attempting an English-Hindi Anusaaraka machine translation system.

Anusaaraka mainly focus on language access between Indian languages, using principles of Paninian Grammar (PG) (Bharati *et al*., 1995) and exploiting the close similarity of Indian languages.

Mantra machine translation system: MAchiNe assisted TRAnslation tool (MANTRA) (1999). It translates English text into Hindi in a precise domain of personal administration, specifically gazette notifications, office orders, office memorandums and circulars. Initially, the Mantra system was started with the translation of administrative document such as appointment letters, notification and circular issued in central government from English to Hindi. It is based on the Tree Adjoining Grammar (TAG) formalism from University of Pennsylvania. It uses Lexicalized Tree Adjoining Grammar (LTAG) (Bandyopadhyay, 2004) to represent the English as well as the Hindi grammar. Tree Adjoining Grammar (TAG) uses for parsing and generation.

It is based on synchronous Tree Adjoining Grammar and uses tree transfer for translating from English to Hindi. The system is tailored to deal with its narrow subject domain. 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.

This system can be obtained from the C-DAC website (http://cdac.in/html/aai/mantra.asp). About this system the contact person is Dr. Hemant Darbari and Dr. Mahendra Kumar Pandey. This project was funded by the Rajya Sabha Secretariat. 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-domain. The system is ready for use in its domain. The system is developed for the RajyaSabha Secretariat, the Upper House of Parliament of India. It translate the proceedings of parliament such as study to be Laid on the Table, Bulletin Part-I and Part-II. This system also works on other language pairs such as English- Bengali, English-Telgu, English-Gujarati and Hindi-English and also among Indian languages such as Hindi-Bengali and Hindi-Marathi. The Mantra approach is general, but the lexicon/grammar has been limited to the sub-language of the domain.

MaTra system: The MaTra system (2004), developed by the Natural Language group of the Knowledge Based Computer Systems (KBCS) division at the National Centre for Software Technology (NCST), Mumbai (currently CDAC, Mumbai) and supported under the TDIL Project is a tool for human aided machine translation from English to Hindi for news stories.

It has a text categorization component at the front, which determines the type of news story (political, terrorism, economic and so on.) before operating on the given story. Depending on the type of news, it uses an appropriate dictionary. It requires considerable human assistance in analyzing the input. Another novel component of the system is that given a complex

English sentence, it breaks it up into simpler sentences, which are then analyzed and used to generate in Hindi. They are using the translation system in a project on Cross Lingual Information Retrieval (CLIR) (Rao, 2001) that enables a person to query the web for documents related to health issues in Hindi.

Mantra machine translation: The English to Hindi Anusaaraka system follows the basic principles (Bharati *et al*., 1997) of information preservation. The system makes text in one Indian language accessible in another Indian language. It uses XTAG based super tagger and light dependency analyzer developed at University of Pennsylvania for performing the analysis of the given English text. It distributes the load on man and machine in novel ways. The system produces several outputs corresponding to a given input. The simplest possible (and the most robust) output is based on the machine taking the load of lexicon and leaving the load of syntax on man. Output based on the most detailed analysis of the English input text, uses a full parser and a bilingual dictionary. The parsing system is based on XTAG (Bandyopadhyay, 2002) (consisting of super tagger and parser) wherein we have modified them for the task at hand. A user may read the output produced after the full analysis, but when he finds that the system has "obviously" gone wrong or failed to produce the output, he can always switch to a simpler output.

AnglaBharti technology: The AnglaBharti project was launched by Sinha *et al*. (2001) at the Indian Institute of Technology; Kanpur in 1991 for Machine aided Translation from English to Indian languages. Professor Sinha *et al*. (2001) has pioneered Machine Translation research in India. The approach and lexicon of the system is general-purpose with provision for domain customization. A machine-aided translation system specifically designed for translating English to Indian languages. English is a SVO language while Indian languages are SOV and are relatively of free word-order. Instead of designing translators for English to each Indian language, AnglaBharti uses a (Dave *et al*., 2001) pseudo-interlingua approach. It analyses English only once and creates an intermediate structure called Pseudo Lingua for Indian Languages (PLIL).

In AnglaBharti they use rule based system with context free grammar like structure for English, A set of rules obtained through corpus analysis which is used to distinguish conceivable constituents. Overall, the AnglaHindi (Sinha and Jain, 2003) system attempts to generalizing the constituents and replacing them with abstracted form from the raw examples. The abstraction integrate example-based approach with rule-based and human engineered post-editing.

AnglaBharti is a pattern directed rule based system with context free grammar (Sinha and Jain, 2003) like structure for English (source language) which generates a 'pseudo-target' (PLIL) applicable to a group of Indian languages (target languages). A set of rules obtained through corpus analysis is used to identify plausible constituents with respect to which movement rules for the PLIL is constructed. The idea of using PLIL is primarily to exploit structural similarity to obtain advantages similar to that of using Interlingua approach. It also uses some example-base to identify noun and verb phrasal's and resolve their ambiguities.

AnglaBharti-II: AnglaBharti-II (2004) (Sinha, 2004) addressed many of the shortcomings of the earlier architecture. It uses a Generalized Example-Base (GEB) for hybridization besides a Raw Example-Base (REB). During the development phase, when it was found that the modification in the rule-base was difficult and might 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 were made for automated pre-editing and paraphrasing,

The purpose of automatic pre-editing module is to transform/paraphrase the input sentence to a form which is more easily translatable. Automated pre-editing may even fragment an input sentence if the fragments are easily translatable and positioned in the final translation Such fragmentation may be triggered by in case of a failure of translation by the 'failure analysis' module. The failure analysis consists of heuristics on speculating what might have gone wrong. The entire system is pipelined with various sub-modules. All these have contributed significantly to greater accuracy and robustness to the system.

Anubharti technology: Anubharti (2004) (Sinha, 2004) approach for machine-aided-translation is a hybridized example-based machine translation approach that is a combination of example-based, corpus-based approaches and some elementary grammatical analysis. The example-based approaches follow human-learning process for storing knowledge from past experiences to use it in future. In Anubharti, the traditional EBMT (Gupta and Chatterjee, 2003) approach has been modified to reduce the requirement of a large example-base. This is done primarily by is achieved by identifying the syntactic groups. Matching of the input sentence with abstracted

examples is done based on the syntactic category and semantic tags of the source language structure.

Both of these system architectures, AnglaBharti and AnuBharti, have undergone a considerable change from their initial conceptualization. In 2004 these systems named as AnglaBharti-II and AnuBharti-II. AnglaBharti-II uses a generalized example-base for hybridization besides a raw example-base.and the 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.

Anuvaadak machine translation: Anuvaadak 5.0 system has been developed by super Info soft private limited, Delhi under the supervision of Mrs. Anjali Rowchoudhury for a general purpose English-Hindi Machine Translation. For specific domains it has inbuilt dictionaries. It has specific domains like Official, formal, agriculture, linguistics, technical and administrative. The meaning of any English word is not available in Hindi in dictionary then there is facility of translation is provided. In the windows family this software runs on any Operating system.

Tamil-Hindi machine aided translation system: The system Tamil-Hindi Machine-Aided Translation system has been developed by Prof. C.N. Krishnan at Anna University at KB Chandrashekhar (AU-KBC) research centre, Chennai. The translation system is based on Anusaaraka Machine Translation System, the input text is in Tamil and the output can be seen in a Hindi text.

It uses a lexical level translation and has 80-85% coverage. Stand-alone, API and Web-based on-line versions are developed. Tamil morphological analyser and Tamil-Hindi bilingual dictionary are the by-products of this system. They also developed a prototype of English-Tamil Machine-Aided Translation system. It includes exhaustive syntactical analysis. It has limited vocabulary (100-150) and small set of transfer rules. The system can be accessed at http://www.au-kbc.org/research-areas/nlp/demo/mat/.

English-Kannada machine-aided translation system: English-Kannada MAT system is developed at Resource Centre for Indian Language Technology Solutions (RC-ILTS), University of Hyderabad by Dr. K. Narayana Murthy. The system is essentially a transfer-based approach and it has been applied to the domain of government circulars. English-Kannada machine translation system using Universal Clause Structure Grammar (UCSG) formalism. The system is funded by the Karnataka government.

UNL-based English-Hindi machine translation system: The Universal Networking Language (UNL) used as Interlingua for English to Hindi translation, it was developed by the Indian Institute of technology, Bombay. Prof. Pushpak Bhattacharyya working on machine translation system from English to Marathi and Bengali using the UNL formalism.

Shiva and Shakti machine translation: The system Shiva is an Example-based and the system Shakti is working for three target languages like Hindi, Marathi and Telgu. Shiva and Shakti are the two Machine Translation systems from English to Hindi and are developed jointly by Carneige Mellon University USA, international institute of information technology, Hyderabad and Indian institute of science, Bangalore, India. The system is used for translating English sentences into the appropriate language. Shakti machine translation system (Bharati *et al*., 2003) has been designed to produce machine translation systems for new languages rapidly. Shakti system combines rule-based approach with statistical approach whereas Shiva is Example-Based machine translation system. The rules are mostly linguistic in nature and the statistical approach tries to infer or use linguistic information. Some modules also use semantic information. Currently system is working for three languages (Hindi, Marathi and Telugu).

Anubaad hybrid machine translation system: Anubaad a hybrid MT system is developed in the year 2004 for translating English news headlines to Bengali, developed by Bandyopadhyay (2000) at Jadavpur University Kolkata and. The current version of the system works at the sentence level.

Hinglish machine translation system: Hinglish a machine translation system for pure (standard) Hindi to pure English forms developed by Sinha and Thakur (2005) in the year 2004. It had been implemented by incorporating additional level to the existing English to Hindi translation (AnglaBharti-II) and Hindi to English translation (AnuBharti-II) systems developed by Sinha. The system claimed to be produced satisfactory acceptable results in more than 90% of the cases. Only in case of polysemous verbs, due to a very shallow grammatical analysis used in the process, the system is not capable to resolve their meaning.

English to (Hindi, Kannada, Tamil) and Kannada to Tamil language-pair example based machine translation system: English to {Hindi, Kannada and Tamil} and Kannada to Tamil language-pair example

based machine translation system developed by Balajapally *et al*. (2006) in the year 2006. It is based on a bilingual dictionary comprising of sentence-dictionary, phrases-dictionary, words-dictionary and phonetic-dictionary is used for the machine translation. Each of the above dictionaries contains parallel corpora of sentence, phrases and words and phonetic mappings of words in their respective files. Example Based Machine Translation (EBMT) has a set of 75,000 sentences most commonly spoken that are originally available in English. These sentences have been manually translated into three of the target Indian languages, namely Hindi, Kannada and Tamil.

Punjabi to Hindi machine translation system: Punjabi to Hindi machine translation system developed by Josan and Lehal (2008) at Punjabi University Patiala in the year 2007. This system is based on direct word-to-word translation approach. This system consists of modules like pre-processing, word-to-word translation using Punjabi-Hindi lexicon, morphological analysis, word sense disambiguation, transliteration and post processing. The system has reported 92.8% accuracy.

Sampark: Machine translation System among Indian language: Sampark: Machine translation system among Indian languages developed by the Consortium of institutions. Consortiums of institutions include IIIT Hyderabad, University of Hyderabad, CDAC (Noida, Pune), Anna University, KBC, Chennai, IIT Kharagpur, IIT Kanpur, IISc Bangalore, IIIT Alahabad, Tamil University, Jadavpur University in the year 2009. Currently experimental systems have been released namely {Punjabi, Urdu, Tamil, Marathi} to Hindi and Tamil-Hindi Machine Translation systems.

Hindi to Punjabi machine translation system: Main-Hindi to Punjabi Machine translation System developed by Goyal and Lehal (2010) at Punjabi University Patiala in the year 2009. This system is based on direct word-to-word translation approach. This system consists of modules like pre-processing, word-to-word translation using Hindi-Punjabi lexicon, morphological analysis, word sense disambiguation, transliteration and post processing. The system has reported 95% accuracy.

The overall conclusion of machine translation systems in Indian perspectives that from the year 1995 to 2009 the MT systems developed have achieved lots of success in translating languages. Still work has been carried out to achieve better than previous study.

Overview of machine translation system: Machine Translation is the application of computers to the translation of texts from one natural language into another natural language. It is an important sub-discipline of the wider field of artificial intelligence. The conclusion of machine translation systems that have been developed in India for translation from English to Indian languages is shown in the Table 1 and 2.

Table 1: The overview of machine translation systems

Systems	Year	Organization/Institute	Team
Anusaaraka	1995	IIT Kanpur	Prof. Rajeev Sangal and Team
Mantra	1999	C-DAC, Bangalore	Dr. Hemant Darbari and Dr. Mahendra Kumar Pandev and Team
Matra	2004	C-DAC, Mumbai	Dr. Durgesh Rao and Team
AnglaBharti	1991	IIT Kanpur	Prof. R.M.K. Sinha and Team
AnuBharti	2004	IIT Kanpur	Prof. R.M.K. Sinha and Team
Shiva and Shakti	2004	Carneige Mellon University USA, IIIT Hyderabad and Institute of Science, Bangalore	Shiva and Shakti Machine Translation Team.
Anubaad	2004	Jadavpur University, Kolkata	Dr. Sivaji Bandyopadhyay and team
Sampark	2009	Consortium of Institutions	Sampark machine translation team

Table 2: Summary of machine translation system

Systems	Conclusions
Anusaaraka	It is domain free but the system has been applied mainly for translating children's stories. It aims for perfect "information preservation". It mainly focuses on language access between Indian languages.
Mantra	The system is developed for the Rajya Sabha Secretariat, the upper house of parliament of India, but now it also works on Indian language pairs.
Matra	It is a text categorization component it breaks up the complex English Sentences into Simpler sentences which then analyzed and used to produce in Hindi.
Angla Bharti	The approach and Lexicon of the system is general Purpose with provision for domain customization.
Hinglish machine	It had been implemented by incorporating additional level to the existing translation system English to Hindi translation The System claimed to be produced satisfactory acceptable results in more than 90% of the cases.
Shiva and Shakti	The Shiva machine translation system is used for translating English sentences into the appropriate language.
Machine Translation System	Shakti machine translation system has been designed to produce machine translation systems for new languages Hindi, Marathi and Telugu.

With the above details and information, an important feature of the MT system on this task is the correct manipulation of the terms and concepts of the domain. The main goal of MT systems is correctly identify and process them with high quality.

CONCLUSION

Machine translation is relatively new in India-about two decades of research and development efforts. the goal of TDIL project and the various resource centres under the TDIL project works on developing machine translation systems for Indian languages. 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 word nets. The NLP association of India, regular international conferences like International National Conference on Natural Language processing (ICON) and lexical resource E groups like (lr_egroup@iiit.ac.in) are consolidating and coordinating NLP and MT efforts in India.

REFERENCES

Balajapally, P., P. Pydimarri, M. Ganapathiraju, N. Balakrishnan and R. Reedy, 2006. Multilingual book reader: Transliteration, word-to-word translation and full-text translation. Proceeding of the 13th Biennial Conference and Exhibition Conference of Victorian Association for Library Automation Melbourne, Feb. 8-10, CMU, Australia, pp: 1-12.

Bandyopadhyay, S., 2000. ANUBAAD-the translator from English to Indian languages. Proceedings of the 7th State Science and Technology Congress, (SSTC'00), Calcutta, India, pp: 1-9.

Bandyopadhyay, S., 2002. Teaching MT: An Indian perspectives. Proceeding of the UMIST 6th EAMT Workshop, Teaching Machine Translation, Nov. 14-15, MT-Archive, Manchester, England, pp: 13-22.

Bandyopadhyay, S., 2004. Use of machine translation in India. AAMT J., 36: 25-31.

Bharati, A., V. Chaitanya, A.P. Kulkarni and R. Sangal, 1997. Anusaaraka: Machine translation in stages. Vivek Q. Artif. Intell., 10: 22-25.

Bharati, A., V. Chaitanya and R. Sangal, 1995. Natural language processing: A painting perspective. Q. Artif. Intell., 10: 22-25.

Bharati, A., R. Moona, P. Reddy, B. Sankar and D.M. Sharma *et al*., 2003. Machine translation: The Shakti approach. Proceeding of the 19th International Conference on Natural Language Processing, Dec. 2003, MT-Archive, India, pp: 1-7.

Dave, S., J. Parikh and P. Bhattacharyya, 2001. Interlingua-based English-Hindi machine translation and language divergence. J. Mach. Trans., 16: 251-304.

Goyal, V. and G.S. Lehal, 2010. Web based Hindi to Punjabi machine translation system. J. Emerg. Technol. Web Intell., 2: 148-151.

Gupta, D. and N. Chatterjee, 2003. Identification of divergence for English to Hindi EBMT. Proceeding 10th of the MT SUMMIT, (SUMMIT'03), MT-Archive, New Orleans, Louisiana, USA., pp: 1-8.

Josan, G.S. and G.S. Lehal, 2008. Punjabi to Hindi machine translation system. Proceedings of the 22nd International Conference on Computational Linguistics, Aug. 21-24, MT-Archive, Manchester, UK., pp: 157-160.

Kumar, G.B. and K.N. Murthy, 2006. UCSG shallow Parser. Lecture Notes Comput. Sci., 3878: 156-167. DOI: 10.1007/11671299

Naskar, S. and S. Bandyopadhyay, 2002. Use of machine translation in India: Current status. Proceeding of the 7th EAMT Workshop on Teaching Machine Translation, (TMT'02), MT-Archive, Manchester, UK., pp: 23-32.

Rao, D., 2001. Machine translation in India: A brief survey. Proceedings of SCALLA 2001 Conference, (SCALLA'01), National Centre for Software Technology. Banglaore, India, pp: 1-6.

Sinha, R.M.K. and A. Thakur, 2005. Machine translation of bi-lingual Hindi-English (Hinglish) text. Proceeding of the 10th Conference on Machine Translation, Sept. 13-15, MT-Archive, Phuket, Thailand, pp: 149-156.

Sinha, R.M.K., R. Jain and A. Jain, 2001. Translation from English to Indian languages: Anglabharti approach. Proceeding of the Symposium on Translation Support System, Feb. 15-17, IIT, Kanpur, India, pp: 15-17.

Sinha, R.M.K. and A. Jain, 2003. Anglahindi: An English to Hindi machine-aided translation system. Proceeding of the 9th MT Summit, (MTS'03), MT-Archive, New Orleans, Sept. 23-27, USA., pp: 494-497.

Sinha, R.M.K., 2004. An engineering perspectives of machine translation: AnglaBharti-II and AnuBharti-II architectures. Proceeding of the International Symposium on Machine Translation, NLP and Translation Support System, Nov. 17-19, Tata McGraw Hill, New Delhi, pp: 1-9.