CURRENT STATUS OF MACHINE TRANSLATION: INDIAN PERSPECTIVE

Abhijit Paul[1], Bipul Syam Purkayastha[2]
[1,2]Department of Computer Science, Assam University Silchar

ABSTRACT
The purpose of this paper is to give a view of the current status of research, development of Machine Translation (MT) in India. Though a lot of work has been done on Machine Translation for Indian languages, but a complete automatic high quality Machine Translation System is still far off; limited success has been achieved within a restricted domain. Currently the machine translation systems that are available in India, work for translation from English to Indian languages, Indian to English language and among Indian languages. Most of them use domains viz. general, government documents/reports and news stories, children stories, etc. This paper is also discussed characteristics of Indian languages and some of the existing Indian Machine Translation Systems along with their features, domains and the approaches used.
Keywords: Machine Translation, Computational Linguistics, Indian languages, Domains

1. INTRODUCTION

Machine Translation (MT) is the field of computational linguistics (CL) and Natural Language Processing (NLP) that explores the use of computer/mobile application to translate text or speech from one natural language known as the source language to another known as target language. At its initial stage, MT performs substitution of words in one natural language to words in other natural language. Current machine translation software often allows users to customize the domain into the MT system, so that the translation result can improve. It shows that current machine translation produces usable output of government and legal documents which are more readily than conversation or less standardized text.

The demand of a Machine Translation System has greatly increased in India due to increasing cross-regional communication across the state and for information exchange. Now a day in India, most material needs to be translated, viz. scientific and technical documentation, any kind of instruction manuals, legal documents, textbooks, publicity leaflets, newspaper reports etc. Due to cross-regional communication nature in India, it is quite challenging and difficult task and it requires consistency and accuracy. Also for professional translators, it is difficult to meet the increasing demands of translation. In such a scenario the machine translation can treat as a substitute.

India is the largest democratic & a multilingual country in the world and there are more than 30 languages and approximately 2000 dialects are used for the communication by the Indian peoples and out of these Hindi and English languages are taken for official work. There are 22 scheduled languages under the eighth schedule of the Indian constitution lists, which are used by the different states for their administrative work and communication purposes. These 22 languages are Assames, Bengali, Bodo, Dogri, Gujrati, Hindi, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Kannada, Kashmiri, Konkani, Maithili, Santali, Sindhi, Tamil, Telugu and Urdu. Because of different multilingual environment in India there is a great demand for

machine translation for the transfer of information and sharing of the ideas among. People of different regions perform their work in their respective regional languages, whereas the Union Government offices work performed their work in English language or Hindi Language. Here the language barrier comes because people of one state can't communicate with the union government or people of another state through their own language. Also in India because of different culture there are different newspapers published locally as well as globally. So to minimize the gap and synchronize between the state government and the central / Union government there is a great demand for translation from regional languages to English language and vice versa. So from the above discussion and survey it is clear that there is a large scope of translation of text from English to Indian Languages and vice versa.

2. CHARACTERISTICS OF INDIAN LANGUAGES

A wide variety of languages are used in India. In this section some of the important characteristics of Indian languages are discussed. Indian languages can be categorized into following four broad families **[1]**:-

1. Indo-Aryan (e.g. - Hindi, Nepali, Bangla, Asomiya , Punjabi, Marathi, Gujrati, Oriya etc).
2. Dravidian (Tamil, Telugu, Kannada, Malayalam etc)
3. Austro-Asian
4. Tibetian-Burmese

Languages within each group are structurally similar to each other, to a great extent. We summarize common characteristics of Indian languages considering their common uses -

i. Indian languages have SOV (Subject – Object – Verb) as the default sentence structure.

ii. Indian languages are free order, i.e. words can be moved freely within a sentence without changing the meaning of the sentence.

iii. Indian languages have a relatively rich set of morphological variants. Unlike English, Indian languages adjectives undergo morphological changes depending upon number and gender.

iv. Indian languages make extensive and productive use of complex predicates (CPs). A complex predicate combines a light verb with a verb, noun or adjective to produce a new verb, e.g. "Savita *aa gayi"* which means "Savita arrived".

v. Indian languages make use of post-position case markers (Karaks) instead of prepositions.

vi. Most Indian languages have only two genders – masculine and feminine.

vii. Some adjectives are also modified to agree with gender e.g. - achcha ladka vs achchi ladki.

viii. Unlike English, Indian-language pronouns have no associated gender information

3. TRANSLATION INVOLVING INDIAN LANGUAGES

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 **[2]:**-

i. The research and development projects at Indian Institute of Technology (IIT), Kanpur

ii. The research and development projects at Indian Institute of Technology (IIT), Bombay

iii. National Centre for Software Technology (NCST) Mumbai (now, Centre for Development of Advanced Computing (CDAC), Mumbai

iv. Computer and information Sciences Department, University of Hyderabad

v. Centre for Development of Advanced Computing (CDAC), Pune

vi. Ministry of Communications and Information Technology

vii. Government of India, through its Technology Development in Indian Languages (TDIL) Project.

Following are the Machine Translation Systems available for English to Indian, Indian to English and Indian to Indian language.

3.1 Development of Indian to Indian languages Machine Translation System

Anusaaraka systems among Indian languages: The Anusaaraka MT project was started in 1997 at IIT Kanpur, by Prof. Rajeev Sangal. And now it is being continued at IIIT Hyderabad. It is mainly concerned with the explicit aim of translation among Indian languages. This project was funded by the Technology Development in Indian Languages (TDIL). It has been built from Bengali, Telugu, Punjabi, Kannada and Marathi to Hindi. It is a domain free system, but the system mainly applied to translating children's stories. It

mainly uses Paninian Grammar (PG) for translation between Indian languages [3].

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

Hindi to Punjabi machine translation system: Hindi to Punjabi MT System developed at Punjabi University, Patiala in the year 2010 by Goyal and Lehal. Direct word-to-word translation approach has been used in this MT system. This system consists of different important modules for MT like pre-processing, word-to-word translation using Hindi-Punjabi lexicon, morphological analysis, transliteration, word sense disambiguation and post processing. The system achieved accuracy over 95% **[4].**

Punjabi to Hindi machine translation system: In the year 2007, this translation system was developed by Josan and Lehal at Punjabi University, Patiala. This systems works in reverse order of the Hindi to Punjabi MT system. It also follows the direct word- to-word translation approach and consists of the same modules like Hindi to Punjabi MT system. The system has reported accuracy over 92% **[5].**

Kannada to Tamil language-pair example based machine translation system: Kannada to Tamil machine translation system developed by Balajapally and others in the year 2006. Language pair example based approach has been used for machine translation. Example based in terms of bilingual dictionary, parallel corpus, phrases and words and phonetic mappings of words are used **[6]**.

Tamil-Hindi machine aided translation system: The system Tamil to Hindi Machine-Aided Translation system has been developed at Anna University, Chennai. The translation system is based on MT project Anusaaraka. In the system the input text is in Tamil and the output can be seen in a Hindi text.

Telugu-Tamil MT System: The system developed by **[7]** and it uses the Telugu Morphological analyzer and Tamil generator for translation. The system also uses the Telugu-Tamil dictionary developed as a part of MAT Lexica.

Bengali to Hindi MT System: The hybrid based approach has used in this system. Hybrid based approach consists of statistical and lexical transfer based techniques. The system shows the BLEU scores of SMT and lexical transfer based system are 0.1745 and .0424 respectively. The performance of the hybrid system is better and its BLEU score is 0.2275 **[8].**

Lattice Based Lexical Transfer in Bengali Hindi MT Framework: A Bengali to Hindi MT framework has been designed to transfer lexicon and in this system transfer based MT approach has been used. Mainly the system replaced Bengali word with most frequent Hindi word. However, the most frequent translation may not be correct if it is considered the context of the word(s) **[9].**

Bengali to Assamese MT System: Example Based Machine Translation (EBMT) approach has been used in this system and the system works Bengali-Assamese News domain. It is also includes some preprocessing and post-editing modules **[10].**

Sanskrit-Hindi Anusaarka [11]: Development of Sanskrit to Hindi MT system was proposed in 2009 under the Anusaarka project and it allows the user to access the source language text and give the rough output in the target language.

Tamil-Malayalam MT System: The system developed at Bharathidasan University, Tamilnadu in 2007 and is working on translating between languages belonging to the same family, such as Tamil-Malayalam translation. This system includes different modules like Lexical database, Suffix database, Morphological Analyzer, Syntactic Analyzer.

3.2 Development of Indian to English languages Machine Translation System

Hinglish machine translation system: Hinglish (Hindi to English) machine translation system developed by Sinha and Thakur in the year 2004. The system mainly an extension of the existing English to Hindi translation (AnglaBharti-II) and Hindi to English translation (AnuBharti-II) systems, which was also developed by Sinha. The system achieves accuracy over 90%. In case of polysemous words & verbs, the system is not capable to resolve their meaning because of its very shallow grammatical analysis **[12]**.

3.3 Development of English to Indian languages Machine Translation System

Mantra machine translation system: MAchiNe assisted TRAnslation (MANTRA) tool was developed in 1999. Initially, the system was started with the translation of administrative document such as circular issued in central government, appointment letters, notification etc. from English to Hindi. Now it translates English text into Hindi in a precise domain of personal administration, gazette notifications, office memorandums & orders, and circulars. It is based on Lexicalized Tree Adjoining Grammar (LTAG) and for translating from English to Hindi, it uses tree transfer. MANTRA was funded by the Rajya Sabha Secretariat. This system also works on other language pairs such as English- Bengali, English-Telgu, English- Gujarati and Hindi-English **[13]**.

MaTra system: The MaTra system was developed in 2004. It has been developed by the Natural Language Processing group at the National Centre for Software Technology (NCST), Mumbai (currently CDAC, Mumbai). The system supported under the TDIL Project for translation from English to Hindi news stories (politics, terrorism, economic and so on). It follows a rule based approach and depending on the type of news, it uses an appropriate dictionary **[14].**

AnglaBharti technology: The AnglaBharti project was developed at the Indian Institute of Technology (IIT), Kanpur in 2001. The system works for translation from English to Indian languages. Instead of writing rules for each Indian language the system used Interlingua transfer based approach. Along with the rules, the system uses some examples to identify noun and verb phrasal's to resolve ambiguities **[15].**

AnglaBharti-II: AnglaBharti-II was developed in 2004. The system addressed many of the loopholes of the earlier architecture, which are used for machine translation. Since the rule based approach was difficult and produces unpredictable result, so the system follows example based approach **[16]**.

Shakti (2003) : The system translates English to any Indian languages. It follows both rule-based approach and statistical approach. Sixty nine modules are there in this system. Out of sixty nine modules, nine modules are used for analyzing the source language, i.e. English, twenty four modules are used for performing bilingual tasks, and the remaining modules are used for generating target Indian language **[17].**

OMTrans(2004)
English to Oriya machine translation system was developed by Mohanty & Balabantaray. Object oriented concept was used in this system. Word Sense Disambiguation (WSD) is also handled in this system **[18]**.

An **English–Hindi Translation System** developed in the year 2002. The system follows transfer based translation approach, which uses different grammatical rules of source and target languages and a bilingual dictionary for translation. Mainly the system works for domain of weather news. The system consists of different modules like pre-processing, English tree generator, post- processing of English trees, generation of Hindi tree, Post-processing of Hindi tree and generating output **[19].**

English to (Hindi, Kannada, Tamil) language-pair example based machine translation system: English to {Hindi, Kannada and Tamil} language-pair example based machine translation system developed by in the year 2006. Language pair example based approach means bilingual dictionary, parallel

corpora, phrases and words and phonetic mappings of words are used.

Anubaad hybrid machine translation system: Anubaad MT system was developed at Jadavpur University, Kolkata in the year 2004. It is translating news stories from English to Bengali. Currently the current version of the system works at the sentence level **[20]**.

UNL-based English-Hindi machine translation system: This system was developed at Indian Institute of technology, Bombay by Prof. Pushpak Bhattacharyya. The system used as Interlingua for English to Hindi translation. It also works for English to Marathi and Bengali.

Anuvaadak machine translation: Anuvaadak system has been developed for translating Official, linguistics, formal, technical, agricultural, and administrative documents from English to Hindi. It has been developed by super Info soft private limited, Delhi under the supervision of Mrs. Anjali Rowchoudhury.

English-Kannada machine-aided translation system: English-Kannada machine translation system was developed at the University of Hyderabad by Dr. K. Narayana Murthy. It uses an interlingual transfer based approach and the system mainly works for the domain of government circulars. The system was funded by the Karnataka government.

English to Indian Languages MT System (E-ILMT) (2006)

The EILMT System works for English to Indian Languages machine translation. It works for Tourism and Healthcare Domains. It is developed by a Consortium of Nine institutions, namely C-DAC MumbaiIIIT Hyderabad, C-DAC Pune, IIT Mumbai, Jadavpur University Kolkata, IIIT Allahabad, Utkal University Bangalore, Amrita University Coimbatore and Banasthali Vidyapeeth, Banasthali. The project is funded by Department of Information Technology, MCIT Government of India **[21].**

4. CONCLUSION

Though the machine translation is the earliest applications of NLP and so many MT systems are available in India for translating text from Indian to English, English to Indian and Indian to Indian languages, but the goal of achieving error-free translation that reads fluently in target languages is still far off; limited success has been achieved within a restricted domain. Also the MT systems that are available in India for Indian languages mainly work for highly spoken languages like Hindi, Bengali, Marathi, Gujarati, Tamil, Telegu, Oriya, Punjabi, Assamese, Kannada, Malayalam, etc., but due to limited resources, development of other languages (Nepali, Bodo, Khashi, Mizo, etc.) MT system is still far off.. So the correct MT systems for highly spoken languages as well as least spoken languages MT system still a great demand in India.

REFERENCES

[1] Siddiqui T, and Tiwary U.S, 2010 "Natural Language Processing and Information Retrieval", Oxford University Press Publication, 2010.

[2] Dwivedi S.K. and Sukhadeve P.P, 2010 "Machine Translation System in Indian Perspectives", Journal of Computer Science 6 (10): 1111-1116, ISSN 1549-3636, Science Publications.

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

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

[5] 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.

[6] 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.

[7] Parameswari K, Sreenivasulu N.V., Uma Maheshwar Rao G & Christopher M, (2012) "Development of Telugu-Tamil Bidirectional Machine Translation System: A special focus on case divergence", in proceedings of 11th International Tamil Internet conference, pp 180-191

[8] Sanjay Chatterji, Devshri Roy, Sudeshna Sarkar & Anupam Basu, (2009) "A Hybrid Approach for Bengali to Hindi Machine Translation", In proceedings of ICON-2009, 7th International Conference on Natural Language Processing, pp. 83-91.

[9] Sanjay Chatterji, Praveen Sonare, Sudeshna Sarkar & Anupam Basu, (2011) "Lattice Based Lexical Transfer in Bengali Hindi Machine Translation Framework", in Proceedings of ICON-2011: 9th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from ltrc.iiit.ac.in/proceedings/ICON-2011

[10] Kommaluri Vijayanand, Sirajul Islam Choudhury & Pranab Ratna "VAASAANUBAADA - Automatic Machine Translation of Bilingual Bengali-Assamese News Texts", in proceedings of Language Engineering Conference-2002, Hyderabad, India © IEEE Computer Society.

[11] Akashar Bharti, Amba Kulkarni,"Anusaarka: An Accessor cum Machine Translator", Department of Sanskrit Studies, University of Hyderabad, Hyderabad, 2009.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] Bharati, R. Moona, P. Reddy, B. Sankar, D.M. Sharma & R. Sangal, (2003) "Machine Translation: The Shakti Approach", Pre-Conference Tutorial, ICON-2003.

[18] S. Mohanty & R. C. Balabantaray, (2004) "English to Oriya Translation System (OMTrans)" cs.pitt.edu/chang/cpol/c087.pdf

[19] Lata Gore & Nishigandha Patil, (2002) "English to Hindi - Translation System", In proceedings of Symposium on Translation Support Systems. IIT Kanpur. pp. 178-184**.**

[20] 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.

[21] R. Ananthakrishnan, Jayprasad Hegde, Pushpak Bhattacharyya, Ritesh Shah & M. Sasikumar, (2008) "Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation", in proceedings of International Joint Conference on NLP (IJCNLP08), Hyderabad, India