The Journey of Indian Languages: Perpectives on Culture and Society Study of Machine Translation Systems and Techniques for Indian Languages

Article · January 2019

CITATIONS
0

READS
7

2 authors:

Jatin C. Modh
Gujarat Technological University
12 PUBLICATIONS **38** CITATIONS
SEE PROFILE

Jatinderkumar R. Saini
Symbiosis Institute of Computer Studies and Research
235 PUBLICATIONS **1,490** CITATIONS
SEE PROFILE

Study of Machine Translation Systems and Techniques for Indian Languages

Jatin C. Modh
Assistant Professor,
Narmada College of Computer Application, Bharuch, Gujarat, India.
jatinmodh@yahoo.com

Dr. Jatinderkumar R. Saini
Professor & I/C Director,
Narmada College of Computer Application,Bharuch, Gujarat, India.
saini_expert@yahoo.com

Natural languages vary from region to region in India. There are 22 official languages used in India. Due to the advent of information technology, many documents are available in digitized form. So there is a requirement of Machine Translation Systems in various domains like education, health, business and various government agencies. Machine Translation System (MTS) translates the text from source language into the target language. It is a sub-field of Artificial Intelligence. In Machine Translation System (MTS), one natural language gets converted to other language using computational applications with minimal human effort. Besides English and Hindi, 22 scheduled languages are used as a communication medium in India. There have been many attempts in Machine Translation System for Indian languages. Despite that, presently we do not have an effective Machine Translation System. This paper gives a short description of the work done on various Machine Translation Systems and approaches for Indian languages.

Keywords: Computational Linguistics, English, Hindi, Machine Translation System (MTS)

1. Introduction :

India has much diversity in the area of languages. There are 22 official languages in India. Hindi and English are used for the official work of central government. All the official documents are published in Hindi or English or in both the languages by the Indian government. In the states, regional language or Hindi language is used. In the states of India, local newspapers, magazines and books are published in the regional language only. 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. Manual translation of documents is very time consuming and costly. For the exchange of information among states, central government, industry, academia, a good Machine Translation System (MTS) is required.

Machine Translation is the application of Natural Language Processing (NLP). Machine Translation (MT) [9] is defined as the automated process of translating text from one language called source language to other language called target language. Machine Translation System is helpful in various areas like education, science, media, commerce, technology, culture and government bodies. English language is mostly used as a communication language world-wide. Translating regional languages to English language and vice-a-versa is necessary. There is a great demand for the translation of documents from one language to other language. Many researchers are engaged on the different Machine Translation System projects.

Machine Translation is a very complex task. Machine Translation task faces many challenges like referential ambiguities and complex structures in the source and target language, idioms, collocations, synonym metaphors, polysemy, homonymy and lexical and structural mismatch between source and target languages [3].

2. Overview of Machine Translation Approaches :

Several Machine Translation research works are going on nowadays. Various Machine Translation approaches are suggested by the researchers. Overview of main methods is presented here. There are two broad categorization in Machine Translation Systems; specifically, Rule Based and Empirical Based Machine Translation Systems. Hybrid Machine Translation system enjoys the advantage of both Rule Based Machine Translation System and Empirical Based Machine Translation System. Rule Based Machine Translation System is further classified into direct, transfer and interlingua, while Empirical Based Translation System is further categorized into statistical and example based machine translation system.

2.1 Rule Based Machine Translation (RBMT) :

Rule Based Machine Translation is a simple approach of Machine Translation. Grammar rules

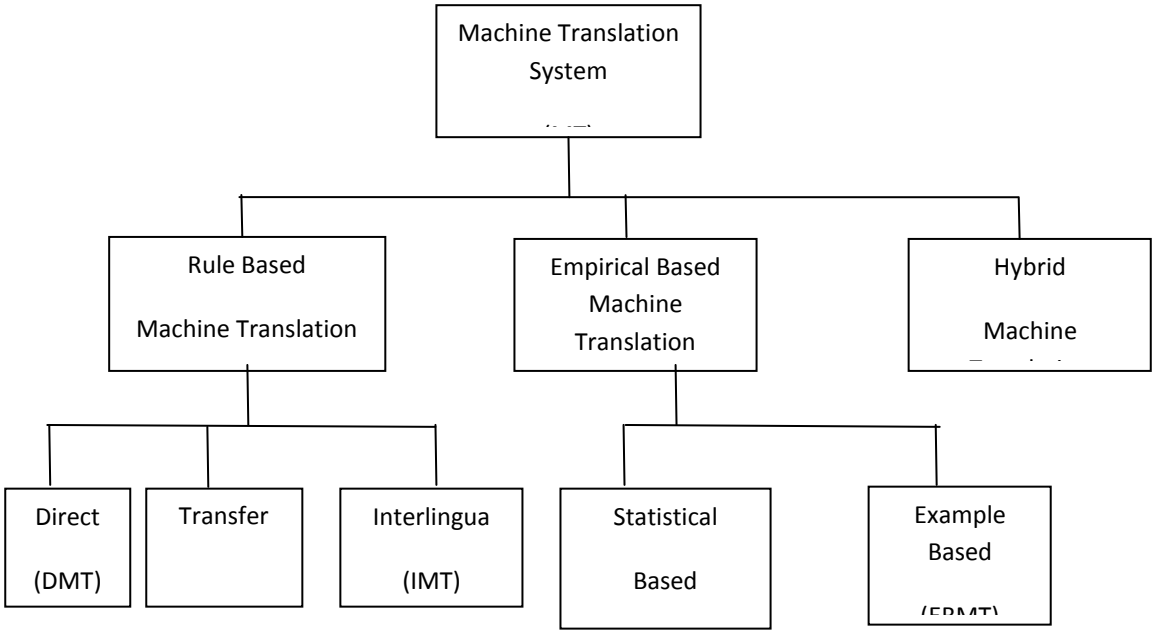

Figure 1: Categories of Machine Translation System

are used by Rule Based Machine Translation. RBMT uses rules of grammar, bilingual lexicon and program to process the rules [3]. RBMT requires human effort to code all of the linguistic resources like source part-of-speech taggers & syntactic parsers, bilingual dictionaries, source to target transliteration, target language morphological generator. Rules play an important role in morphological, syntactic and semantic analysis of each language [6]. It is also known as Knowledge-Based Machine Translation [6]. Dictionary-based i.e. Direct, Transfer-based and Interlingua Machine Translation are the three different methods that come under the RBMT category.

2.1.1 Direct Machine Translation (DMT) :

Direct Machine Translation is a word level approach. Without passing through an intermediary representation, words are translated into target language [6]. In Direct Machine Translation, a different translator is required to build for each pair of source and target language [10]. It is basically bilingual and uni-directional. It is a word-by-word translation approach with little syntactic, semantic analysis and grammatical adjustments. It is less popular method of Machine Translation.

2.1.2 Transfer Based Machine Translation :

There are three stages of Transfer Based Machine Translation; specifically, analysis, transfer and generation. Identification of the components of sentence, structure and parsing of the source language text is performed by the analysis stage. Transformation of source language parse tree into the target language is done by the Transfer stage. Word translation, number, tense representation in the target language is done by the Generation stage. Analysis stage builds a source language dependent representation. For multi-lingual machine translation system, independent transfer component is required for each direction of translation for every pair of languages [3].

2.1.3 Interlingua Based Machine Translation (IMT) :

The main benefit of Interlingua Based Machine Translation is that it can be used with any language pair. Intermediate representation form of the language is used by this method. Interlingua model uses conceptual elements like event, agent, tense etc. Translators take text from input language and then convert it into the intermediate representation. Intermediate representation form is the representation of 'meaning' of input text in some respect. It could form the basis for the generation of target language [6]. This model also builds a parse tree in the language-independent structure, known as Interlingua [29]. For each target language, generator component takes the interlingua as input and generates the translation as output in the target language.

2.2 Empirical Based Machine Translation :

Empirical Based Machine Translation research uses data or corpus based methods. Empirical Based Machine Translation is an alternative method to Rule Based Machine Translation system. Fundamentally Empirical Based Machine Translation techniques use a corpus or database of translated examples [28]. This is the emerging approach that uses large amount of raw data in the form of parallel corpora [28]. Empirical based methods have influenced the Machine Translation research in the present time. Corpus-based approaches are further divided into two sub-approaches, Example Based Machine Translation and Statistical Machine Translation system.

2.2.1 Example Based Machine Translation (EBMT) :

Example Based Machine Translation makes the use of bilingual corpus with parallel texts as its key knowledge and the main objective is translation by similarity. According to Somers [2], three criterias are used for defining EBMT; EBMT uses a bilingual corpus, EBMT uses a bilingual corpus as its main knowledge and EBMT uses bilingual corpus as its main knowledge base, at run-time. It uses a collection of given sentences and corresponding translations with point to point mapping [6]. Translation is done by analogy and example translations are used to train the system.

2.2.2 Statistical Based Machine Translation (SMT) :

Nowadays, Statistical based Machine Translation is the most widely studied Machine Translation approach. It is a Machine Translation approach where translations are made on the basis of statistical models whose parameters are obtained from the analysis of bilingual or multilingual textual corpora [4]. It is a data-oriented approach for converting text from source language to target language. Supervised or unsupervised machine learning algorithms are used for training or learning. Statistical tables are constructed from the corpora with the help of this training [11]. Characteristics of well-formed sentences and correlation between languages are stored in these statistical tables. This information is utilized in decoding

process to find the best translation for the entered text. Statistical Machine Translation is controlled by available parallel corpora. Statistical Machine Translation does not require extensive linguistic analysis. The output is determined by the existing multilingual corpora. Statistical Machine Translation models can comprehend implicit knowledge embodied in co-occurrence statistics [14]. Statistical Machine Translation method faces problems in which deep linguistic knowledge is required.

2.3 Hybrid Machine Translation :

Hybrid Machine Translation approach takes the advantages of both Rule Based Machine Translation and Empirical Based Machine Translation. Several research works use only Rule Based method and several projects use only Statistical Based Machine Translation method. The hybrid approach uses both rules and corpora. The Rule Based Machine Translation systems have high accuracy of linguistic analysis as they can successfully handle word order and overall syntactic structure of the sentences. The Statistical Based Machine Translation systems are built using large collection of previously translated text i.e. parallel text corpora. The Rule Based Machine Translation system has slower development cycle, whereas Statistical Based Machine Translation has faster development cycle. Rule Based Machine Translation systems need manually built lexicon, grammars and algorithms that limit the supported languages. Statistical Based Machine Translation systems are accurate in ambiguity resolution but there is a problem in linguistics. Currently researchers follow Hybrid Machine Translation method in which they integrate multiple Machine Translation approaches within single Machine Translation System [21].

3. Various Machine Translation Systems for Indian Languages :

Machine Translation System should preserve the meaning of original input text. Human translator is still preferred when transferring written text from source language to target language. Many research projects have been carried out in the field of Machine Translation. Various organizations built many Machine Translation systems for Indian languages. The following table represents brief description about Machine Translation Systems involving Indian languages.

Sr No	Name of the System	Developer Organization and Year	Applied Technique(s)	Support for the Indian language(s)	Important Characteristics
1	Google Translator	Franz Josef Och; 2007 [1]	Statistical Based Method	Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Punjabi, Sindhi, Tamil, Telugu and Urdu [5]	It is server-based solution and having Parallel corpora (at least 5,000,000 words /500,000 translation units) [16] ; Having Good enough accuracy [26]
2	Bing Translator	Microsoft; 2009	Statistical Based	Bengali, Hindi and Urdu [7]	The quality of outputs is evaluated using a method called the BLEU (Bilingual Evaluation Understudy) score

					which evaluate the quality of translated text. [27]
3	AnglaMT	Consortium lead by IIT, Kanpur; 2009 [15]	Rule Based	English to Assamese, Bengali, Malayalam, Nepali, Punjabi, Telugu and Urdu	pattern directed rule based system with context free grammar like structure for English (source language)
4	EILMT /ANUVADAKSH	Consortium Based Project; 2009	Hybrid Based [21]	English to Bengali, Bodo, Gujarati, Hindi, Marathi, Oriya, Tamil and Urdu [17]	The training corpus consisted of 12299 sentences and additional 1570 sentences were divided for testing and tuning
5	Sampark Project [12]	Consortium Based Project; 2009	Transfer Based [30]	Tamil to Hindi, Urdu to Hindi, Punjabi to Hindi, Hindi to Punjabi, Telugu to Tamil	Uses Computational Paninian Grammar [8]
6	AnglaBharati-I	IIT, Kanpur; 1991	Pseudo interlingua	English to Indian languages (primarily Hindi)	Machine does 90% translation task where as human does remaining 10% post editing task.
7	AnglaBharati-II	IIT, Kanpur; 2004	Pseudo interlingua [8]	English to Indian languages	The efficiency of the system was improved as compared to ANGLABHARTI-I.
8	MANTRA [13]	CDAC, Pune; 1999	Transfer Based	English to Hindi, Bengali, Telugu, Gujarati; Hindi to English, Marathi, Bengali	Uses Tree Adjoining Grammar (TAG) and Lexicalized TAG.

9	MaTra [13]	CDAC, Mumbai; 2004	Transfer Based	English to Hindi	The 65% of translation done by this system initially was in acceptable state, but it needs to handle the phrasal verbs, interrogative and imperative mood sets.
10	ANUBHARTI-II	2004. IIT, Kanpur	Example Based & Statistic Based	Hindi to other Indian languages [18]	The example based approach emulates human-learning process for storing knowledge from past experiences and to be used in future.
11	HINGLISH	IIT, Kanpur; 2004	Pseudo interlingua	Hindi to English [19]	Produced satisfactory acceptable results in more than 90% of the cases.
12	HINDI-PUNJABI	Vishal & Gurpreet; 2004 [20]	Direct	Hindi to Punjabi	The translation accuracy is 95.40% on the basis of intelligibility test and 87.60% on the basis of accuracy test.
13	SHIVA	IIIT Hyderabad and Carnegie Mellon University USA; 2004 [30]	Example Based	English to Hindi	Easy to extend this system for new target language.
14	SHAKTI	IIIT, Hyderabad and IIS Banglore; 2004 [22]	Rule Based and Statistical approach	English to Hindi, Marathi, Telugu	Consists of 69 different modules, of which 9 modules for analyzing source statements, 24 modules for bilingual tasks, remaining for generating target language.

15	ANUBAAD	S. Bandyopadhyay; 2004 [23]	Example Based [31]	English to Bengali	Shows 98% correct result.
16	VAASAANUBAADA	Vijayanand et. Al; 2002 [24]	Example Based	Bilingual Bengali to Assamese	The bilingual corpus has been constructed.
17	ANUSAARAKA	Akshar Bharti et. Al; 1997 [25]	Direct	Telugu, Kannada, Bengali, Punjabi and Marathi to Hindi	For 80% of Kannada words in dictionary of 30000 root words, Single equivalent word in Hindi.

Table 1: Machine Translation Systems for Indian Languages

4. Conclusion :

Machine Translation is nothing but the automated process that accepts input text in the source language and delivers output text in the target language. Best Machine Translation System keeps the original meaning of the input text as it is. Many words have different meanings in different contexts or References : , so we still require human translator. Based on the study, we found that each method has capability of translation with some limitations. But most of the Machine Translation Systems for Indian languages follow Rule Based, Statistical Based or Hybrid approach. Statistical Based approach brings into maximum accuracy.

It is concluded that currently many researchers are working on Machine Translation Systems involving many Indian languages. Web-based online interactive translation is getting popularity among other methods. Google Translate Machine Translation System supports many Indian languages, which follows Statistical Machine Translation System.

References : :

[1] Wikipedia, Available Online: https://en.wikipedia.org/wiki/Google_Translate

[2] Somers H., Review Article: Example-based Machine Translation, Machine Translation, 14(2), 113-157, 1999

[3] Somya Gupta, "A survey of Data Driven Machine Translation", Department of Computer Science and Engineering Indian Institute of Technology, Bombay, Mumbai, 2012

[4] Statistical machine translation, Available Online: https://en.wikipedia.org/wiki/Statistical_machine_translation, 2017

[5] "Google Translate", Google Corporation Ltd.; Available Online: https://translate.google.co.in/

[6] M. D. Okpor, "Machine Translation Approaches: Issues and Challenges", IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 2, September 2014

[7] "Bing Translator", Microsoft Corporation Ltd.; Available Online: https://www.bing.com/translator

[8] Antony P. J., "Machine Translation Approaches and Survey for Indian Languages", Computational Linguistics and Chinese Language Processing, Vol. 18, No. 1, March 2013, pp. 47-78

[9] J. Hutchins and H. Somers, "An introduction to Machine Translation", Academic Press, 1992

[10] Deepak Khemani, "A First Course in Artificial Intelligence", 2013, pages 702-703

[11] Zhang, Y., "Chinese-English SMT by Parsing", 2006, Available Online: www.cl.cam.ac.uk/~yz360/mscthesis.pdf

[12] Sampark Project, "Sampark: Machine Translation among Indian Languages"; Available Online: http://sampark.iiit.ac.in/sampark/web/index.php/content/demotranslation

[13] Sitender, Seema Bawa, "Survey of Indian Machine Translation Systems", International Journal of Computer Science and Technology (IJCST), March 2012

[14] Sneha Tripathi & Juran Krishna Sarkhel, "Approaches to Machine Translation", International journal of Annals of Library and Information Studies, Vol. 57, 2010, pp. 388-393

[15] AnglaMT, "AnglaMT" Available Online: http://www.cdac.in/index.aspx?id=mc_mat_AnglaMT_mat; Available Online: http://tdil-dc.in/components/com_mtsystem/CommonUI/homeMT.php

[16] Mamta & Tanuj Wala, 2015, "Survey of Approaches Used in Machine Translation System", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 5, May 2015

[17] Anuvadaksh, "Anuvadaksh", Machine Assisted Translation Tool; Available Online: http://eilmt.rb-aai.in/

[18] R.M.K. Sinha,"An Engineering Perspective of Machine Translation", AnglaBharti-II and AnuBharti-II Architectures. In proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS- 2004). November 17-19. Tata McGraw Hill, New Delhi. pp. 134-38, 2004

[19] R. M. K. Sinha, Anil Thakur, "Machine Translation of bi-lingual Hindi-English (Hinglish) text in proceedings of the tenth Machine Translation Summit", MT Summit X, Phuket, Thailand, September 13-15. pp.149-156, 2005

[20] Vishal Goyal, Gurpreet Singh Lehal, "Hindi to Punjabi Machine Translation System", In the Proceedings International Conference for Information Systems for Indian Languages, Patiala, Department of Computer Science, Punjabi University, Patiala, March 9-11, 2011, pp. 236-241, Springer CCIS 139, Germany (2011)

[21] Neeha Ashraf & Manzoor Ahmad, "Experimental Framework using Web-based Tools for Evaluation of Machine Translation Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 4, April 2016, Available online : www.ijarcsse.com

[22] 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

[23] S. Bandyopadhyay, "ANUBAAD - The Translator from English to Indian Languages", In proceedings of the VIIth State Science and Technology Congress. Calcutta. India. pp. 43-51, 2004

[24] Vijayanand K., Choudhury S.I., P., Ratna P., "VAASAANUBAADA - Automatic Machine Translation of Bilingual Bengali-Assamese News Texts", Language Engineering Conference, 2002, Hyderabad, India, pp.183-188

[25] Akshar Bharti, Chaitanya Vineet, Amba P. Kulkarni & Rajiv Sangal, "ANUSAARAKA: Machine Translation in stages", Vivek, a quarterly in Artificial Intelligence, Vol. 10, No. 3, NCST Mumbai, 1997, pp. 22-25

[26] Och, Franz Josef, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Association for Computational Linguistics, pp. 858-867, June 2007

[27] Microsoft Translator, https://en.wikipedia.org/wiki/Microsoft_Translator, 2017

[28] Machine Translation Overview, Available Online: http://language.worldofcomputing.net/machine-translation/machine-translation-overview.html, 2017

[29] Shachi Dave, Jignashu Parikh, and Pushpak Bhattacharyya, "Interlingua based English Hindi machine translation and language divergence", Journal of Machine Translation, 2002

[30] Sugata Sanyal, Rajdeep Borgohain, "Machine Translation Systems in India", Annals of Faculty Engineering Hunedoara, International Journal of Engineering, Tome XI (Year 2013). Fascicule 4. ISSN 1584 – 2673

[31] Goraksh V. Garje, Gajanan K. Kharate, "Survey of Machine Translation Systems in India", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.4, October 2013