Scientific Journal Impact Factor (SJIF): 1.711

e-ISSN: 2349-9745
p-ISSN: 2393-8161

International Journal of Modern Trends in Engineering and Research
www.ijmter.com

Survey Of Machine Translation Development For Indian Regional Languages

Amruta Godase[1], Sharvari Govilkar[2]
[1]*Information Technology (AI & Robotics), PIIT, Mumbai University*
[2]*Computer Engineering, PIIT, Mumbai University*

Abstract— Natural language processing (NLP) is an emerging field of machine learning. NLP deals with many applications which making the use of machine to translate text/speech which is generally termed as "Machine Translation", which is responsible for production of translation from one natural language into another language with or without human support. As India is a multilingual and multicultural country where spoken languages changes after every 50 miles. This necessities the automated machine translation so as to exchange the information and for communication purpose. This paper focuses on different approaches used in the development of machine translation system and also briefly describes existing machine translation systems according to regional languages of India with their features, limitations and domain.

Keywords—Machine translation, computational linguistics, Indian Languages, Indo-Aryan, Rule-based, Statistical, Empirical MT, Principle-based, Knowledge-based, Hybrid

I. INTRODUCTION

India is a multilingual and multicultural country; there are 22 official languages and approximately 2000 dialects which are spoken by different community. Indian languages are divided into five categories: Indo-Aryan (76.87% speakers), Dravidian (20.82% speakers), Austro-Asiatic (1.11% speakers), Tibeto-Burman (1% speakers) and Andmanese (0.001% speakers). English and Hindi are used for official work in most states of India. Hindi and Marathi are the world's fourth mostly commonly used/spoken languages in State of Maharashtra after English. With the advent of Information Technology many documents and web pages are coming up in a local language and also many newspapers are also published in regional languages. Manual translation of these documents is very time consuming and costly. Hence there is need to develop good Machine Translation (henceforth referred as MT) systems to address all these issues in order to establish a better communication between states and union governments to exchange information amongst the people of different states.

Many researchers, Institutions and organizations in India have started working on MT systems for Indian languages and have gained satisfactory results. The research scenario in India is relatively young and MT gained momentum in India with institutions like IIT Kanpur, IIT Mumbai, IIIT Hyderabad, University of Hyderabad and CDAC Pune play a major role in developing these systems. Many MT systems have been developed in India which has used different approaches. This paper provides the brief information about development year, source & target languages, translation approach, domain, salient features and translation accuracy of MT systems in India.

This paper is organized into 5 sections. Section2 gives an introduction of MT, Section 3 discuss major MT systems in India based on language with their features, domain etc. in tabular format; section 4 gives an idea of the different approaches to build a MT systems and finally we conclude the paper in the next section.

II. MACHINE TRANSLATION

Machine translation, an integral part of Natural Language Processing, is important for breaking the language barrier and facilitating the inter-lingual communication where translation is done from source language to target language preserving the meaning of sentence. If we succeed to this, then and only then we can say that exact translation is done by system. MT system holds tremendous potentials in various domains like education, health, business, governmental agencies and information technology.

So we can say for understanding and making communication easy there is a need of MT. This translation can be done by humans, then why there is a need of machine translation? Various reasons are summarised as follows: The first reason is that word of text is huge, hence there is too much that need to be translated. A second reason is that technical materials are too boring for human translator. Thirdly terminology is used consistently in large corporation; they want terms to be translated in the same way every time. Computers are consistent but human do not like to repeat the same translation and fourth reason is that machine-based translation can increase the volume & speed of translation.

III. LITERATURE SURVEY

In this section we now look at some major Indian MT project in details. The parameters we look at are: language pair(s), formalism and strategy for handling problems, ambiguity, complexity & application domain of each MT system. The scope of this paper is restricted to Hindi, Punjabi, Bengali, English and Marathi languages as source/target language.

Table 1. Machine Translation for Hindi language as a Source or Target language

SR No	Machine Translation Systems	Year	Languages for translation	Domain	Approach used	Observations
1.	ANUSAARAK MT [1]	1995	IL-IL	For translating children's stories	Direct based	The focus is not only on MT but also on language access between Indian languages. Currently attempting English-Hindi MT. works on the principles of Paninian Grammar (PG).
2.	MANTRA MT [2][22]	1997	English-Hindi	General	Transfer based	It uses TAG & XTAG. Uses tagger and light dependency analyzer for performing the analysis of I/P English text. It distributes a load on man and machine in novel way.
3.	MANTRA RAJYASABHA [22]	1999	English-Hindi	Office administration documents	Transfer based	System uses TAG & LTAG to represent a grammar. Can preserve the formatting of input word document. Currently working on Hindi to English and Hindi-Bengali.
4.	ANUBHARTI-I [2][11]	2003	Hindi-English	General	Hybrid	Combination of example based, corpus based & some grammatical analysis. It reduces the

						requirement of large example base and it depends on target language.
5.	ANUBHARTI-II [2][11]	2004	Hindi-English	General	Hybrid	It emulates human-learning process for storing knowledge from past experience to use it in future. Shallow chunker is used for fragmentation of input sentences.
6.	Hinglish MT System [26]	2004	Hindi-English	General	Example based	Based on Anubharti-II & Anglabharti-II. It produces satisfactory results in more than 90 cases. It performs shallow grammatical analysis.
7.	An English–Hindi Translation System [2]	2002	English-Hindi	Weather narration	Transfer based	Translation modules consist of preprocessing and post processing of English tree. Also include generation of Hindi tree.
8.	UNL-based English-Hindi MT System [3]	2001	English-Hindi	General	Interlingua	Based on UNL grammar. Easy to add new language for translation.
9.	MaTra: A Practical Approach to Fully-Automatic Indicative English-Hindi Machine Translation [4]	2004-2006	English-Hindi	News, annual report, technical phrases	Transfer based	Based on MSIR. It uses transfer frame like structure representation & also uses heuristics to resolved ambiguities.
10.	A Pure EBMT Approach for English to Hindi Sentence Translation System [5]	2014	English-Hindi	comparing sentence to extract the translation	Example based	This system uses parallel corpora for translating purpose. It contains various modules such as similarity matrix, training matrix & tagging matrix.

Table 2. Machine Translation for Punjabi language as a Source or Target language

SR No	**Machine Translation Systems**	**Year**	**Languages for translation**	**Domain**	**Approach used**	**Observations**
1.	Punjabi to Hindi MT System [6]	2007-2008	Punjabi-Hindi	General	Direct based	Created on windows platform. Requires post-processing. Also contains NER, Word mapping, Ambiguity Resolution. Accuracy is up to 90.67%. WER is 2.34% & SER is 24.26%

2.	Web based Hindi-to-Punjabi MT System [7]	2010	Hindi-Punjabi	Web–pages, email	Direct based	System can translate any complex sentence & system accuracy is up to 95%. Lexicon contains two part one with No disambiguation and second contains multiple meaning depending on context.
3.	Hindi-to-Punjabi MT System [8]	2009-2011	Hindi-Punjabi	General	Direct based	Lookup algorithm and pattern matching algorithms are used. Currently working on Hindi-English. Accuracy is 95.4%.WER is 4.58% & SER is 28.82%. BLUE score is 0.7801

Table 3. Machine Translation for Kannada language as a Source or Target language

SR No	Machine Translation Systems	Year	Languages for translation	Domain	Approach used	Observations
1.	MAT [9][25]	2002	English-Kannada	Government Circulars	Transfer based	Based on UCSG. 40-60% fully accuracy. Post editing tool is provided which outputs the number, type & inter-relationships amongst various clauses in sentences. For each word suitable target equivalence is obtained from bilingual dictionary.

Table 4. Machine Translation for English to Indian language

SR No	Machine Translation Systems	Year	Languages for translation	Domain	Approach used	Observations
1.	ANGLABHARTI-I [2]	2001	English-IL	Public health	Interlingua	Creates a PLIL intermediate structure. The effort of PLIL is 70% and text generation is 30%. Only with 30% new system can be built. In this 90% translation work is done by machine & 10% left to the human post-editing.
2.	ANGLABHARTI-II [2]	2004	English-IL	General	Example based	Provides provisions for automated pre-editing & paraphrasing, conditional multiword expressions as

						well as recognition of named-entities. Contains module for an error analysis, statistical language module.
3.	Shakti [24]	2003	English-IL	General	Transfer based	Linguistic rule based with statistical processing. Consist of various modules for analyzing the source languages, performing the bilingual task and generating target Indian language.
4.	Shiva and Shakti MT System [23][24]	2003	IL-IL	General	Example based	Easy to extend this system for new target language. Rules uses are mostly linguistic in nature. Semantic information is also used by some module.
5.	AnglaHindi [10]	2003	English-Hindi	General	Interlingua	Pseudo interlingua based. Uses all modules of Anglabharti. Makes use of an abstracted example base. Accuracy is 90%.
6.	"English to Indian Languages MT System (E-ILMT)" [11]	2006	English-IL	Tourism and healthcare	Statistical based	The engine was developed using statistical techniques and tools such as fnTBL, Bikel, Pharaoh. Pre-processing phase was included to take care of syntactic re-ordering on the source language to reduce long distance movements through SMT. The syntactically processed corpus was morphologically processed and used for training to tackle the problem of degradation in translation quality. A rule based suffix separation approach was used to separate the root word and the affixes.

Table 5. Machine Translation for Bengali language as a Source or Target language

SR No	Machine Translation Systems	Year	Languages for translation	Domain	Approach used	Observations
1.	ANUBAAD [12]	2000-2004	English-Bengali	News headlines	Example based	Bengali headline is generated after appropriate synthesis if the headline is found in Generalized Tagged Example-base. If the headline cannot be

						translated using Example-base, Generalized Tagged example-base or Phrasal example-base is used then the heuristic translation strategy is used.
2.	VAASAANUBAADA [13]	2002	Bengali-Assamese	News text	Example based	Bilingual corpus is constructed and aligned manually. Longer sentences are fragmented at punctuation to obtain better quality translation.
3.	Exploiting Alignment Techniques in MATREX: the DCU Machine Translation System [14]	2008	English-Bengali	Conference papers	Example based	Makes use of marker-based chunking, which is based on the Marker Hypothesis, psycholinguistic constraint which signifies context. System makes use of an "edit distance style" dynamic programming alignment algorithm.
4.	Bengali to Hindi MT System [15]	2009	Bengali-Hindi	General	Hybrid	Multi-engine Machine Translation approach which Uses an integration of SMT with a lexical transfer based system (RBMT). The BLEU scores of SMT and lexical transfer based system separately are 0.1745 and .0424 respectively The BLEU score of hybrid system is better and it is 0.2275
5.	Lattice Based Lexical Transfer in Bengali Hindi MT Framework [27]	2011	Bengali-Hindi	General	Hybrid	Lattice based integrated with transfer based Uses a lattice-based data structure i.e. a weighted directed acyclic graph with one start node and one end node.

Table 6. Machine Translation for Marathi language as a Source or Target language

SR No	Machine Translation Systems	Year	Languages for translation	Domain	Approach used	Observations
1.	English to Devnagari Translation for UI Labels of Commercial web based Interactive Applications [16]	2011	English-IL	Web based Applications	Hybrid	Used banking glossary available on the web site of RBI to create multilingual dictionary. For lexical analyzer rules are written by C

						languages. Used Bison tools for running system.
2.	Extending capabilities of English to Marathi Machine Translator [17]	2012	English-Marathi	General	Rule-based	Much functionality can be added for improving the performance of translation. It can be expanded by including spelling and grammatical checks, sentiment analysis modules.
3.	Rule based English to Marathi translation of Assertive sentence [19]	2013	English-Marathi	General	Rule-based	Database of set of rules maintained for mapping. Bilingual-Dictionary database plays very important role which is endless. Open-nlp tools performing different processes.
4.	A novel approach for Interlingual example-based translation of English to Marathi [20]	2014	English-Marathi	General	Hybrid	System is trained from bilingual parallel corpora. Sentence pairs contain sentence in one language with their translation into another. Uses parsing techniques.
5.	Transmuter: An approach to rule based English-Marathi machine Translation [21]	2014	English-Marathi	General	Rule-based	The focus is on grammar structure of target language that produces better & smoother translation. Lexicon is built for morphological & semantic properties.

IV. CLASSIFICATION OF MACHINE TRANSLATION

Generally, MT is classified into various categories: rule-based, statistical-based, hybrid-based, example-based, knowledge-based, principle-based, and online interactive based methods. At present, most of the MT related research is based on Rule-based and Hybrid-based approaches. Following figure shows the classification of MT in Natural language Processing (NLP).

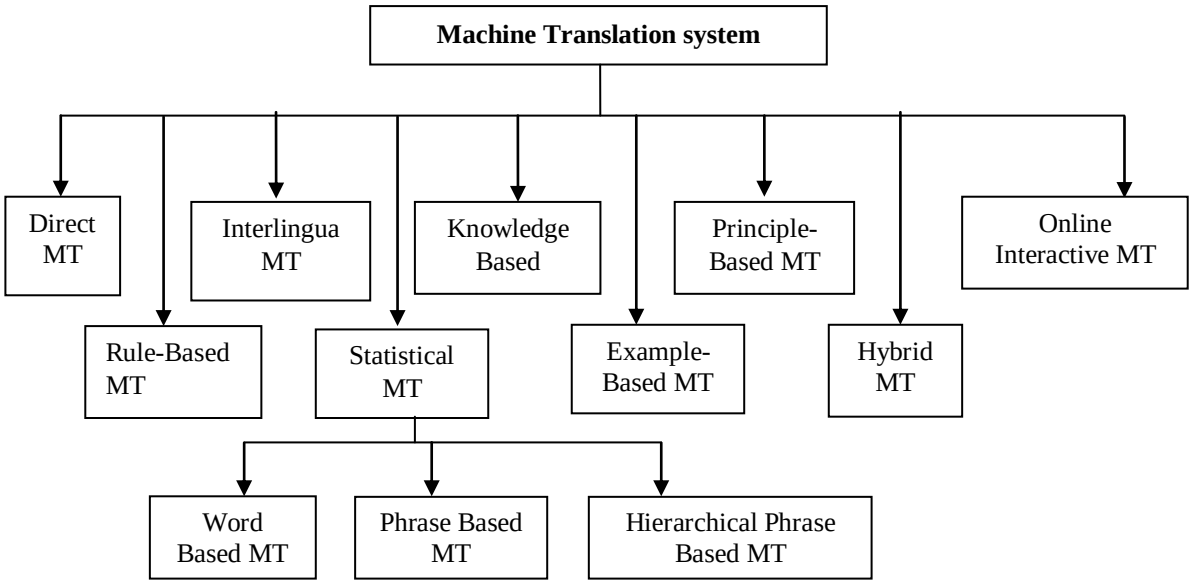

Figure 1.Classifications of MT

1. Direct Translation

Direct Machine Translation is the one of the simplest machine translation approach in which a direct word by word translation of the input source is carried out with the help of a bilingual dictionary and after which some syntactical rearrangement are made for getting correct translation. Typically, the approach is unidirectional and this takes only one language pair into consideration at a time.

2. Rule Based Translation

A Rule-Based Machine Translation (RBMT) system consists of collection of various rules, called grammar rules, a bilingual or multilingual lexicon, and software programs to process the rules. On the basis of the structural differences between the source and target language, a transfer system can be broken down into three different stages: i) Analysis, ii) Transfer and iii) Generation. In the first stage, the parser is used to produce the syntactic representation of a sentence. In the next stage, the result of the first stage is converted into equivalent TL-oriented representations. In the final step TL morphological analyzer is used to generate the final TL texts. Nevertheless, a RBMT system always is extensible and maintainable.

3. Interlingua Based Translation

The next stage of progress in the development of MT systems is the Interlingua approach which intends to translate source language text into more than one language. In this approach, the translation consists of two stages, where the SL is first converted in to the Interlingua (IL) form before translation from the IL to the TL. The main advantage of Interlingua approach is that the analyzer, parser of SL is independent of the generator for the TL and this requires complete resolution of ambiguity in source language text.

4. Statistical-based Approach

Statistical machine translation (SMT) is a data-oriented statistical framework which is based on the knowledge and statistical models which are extracted from bilingual corpora. In this MT, bilingual or multilingual textual corpora of the source/target languages are required. A supervised or unsupervised statistical machine learning algorithm is used to build statistical tables. The collected statistical information is used to find the best translation. There are three different statistical approaches in MT, Word-based Translation, Phrase-based Translation, and Hierarchical phrase based model. In word based translation, the words are translated word by word individually and finally they are arranged in a specific way to get the correct translation. In phrase-based, source and target sentences are divided into phrases. The alignment between words follows certain patterns which are based on flat reordering pattern model. Hierarchical Model is more sophisticated approach which has recursive structures instead of simple phrases.

In SMT, a document is translated according to the probability distribution function which is indicated by $p(e|f)$, which is nothing but the Probability of translating a sentence f in the SL F to a sentence e in the TL E. Finding the best translation is done by picking the one which gives the highest probability, as shown in Equation 1.

$$e = \arg\max p(e \mid f) = \arg\max p(f \mid e)\, p(e) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

5. Example-based translation

The example-based approach relies on large parallel aligned corpora which come under Empirical Machine Translation (EMT). Basic idea of this MT is to reuse the examples of already existing

translations. At run time, an example-based translation is characterized by its use of a bilingual corpus as its main knowledge base. Example-based translation is essentially translation by analogy.

6. Knowledge-Based MT

Knowledge-Based Machine Translation (KBMT) requires functionally complete understanding of the source text prior to the translation into the target text. KBMT is implemented on the Interlingua architecture. Therefore, KBMT must be supported by world knowledge and by linguistic semantic knowledge about meanings of words and their combinations. Thus, in KBMT a specific language is needed to represent the meaning of languages. Once the SL is analyzed, it will run through the augmenter which is the knowledgebase that converts the source representation into correct target representation before synthesizing into the target sentence.

7. Principle-Based MT

Principle-Based Machine Translation (PBMT) Systems are based on the Principles & Parameters Theory of Chomsky‘s Generative Grammar and which employs parsing method. In this, the parser generates a detailed syntactic structure which contains lexical, phrasal, grammatical, and thematic information. This system also focuses on robustness, language-neutral representations, and deep linguistic analyses.

8. Online Interactive Systems

In this online interactive translation system, the user has authority to give suggestion for the correct translation to the translator online. This approach is very useful in a situation where the context of a word is unclear and where many/multiple possible meanings for a particular word. In such cases, the structural ambiguity can be solved with the interpretation of the user.

9. Hybrid-based Translation

By taking the advantage of statistical MT and rule-based MT methodologies, a new approach was developed, which is called “hybrid-based approach”. This approach has better efficiency in the area of MT systems. The hybrid approach can be used in a number of different ways.
1. Translations are performed in the first stage using a rule-based approach which is followed by adjusting or correcting the output using statistical information.
2. Second way in which rules are used to pre-process the input data and for post-process the statistical output of a statistical-based translation system.

V. CONCLUSION

MT is relatively new in India about a decade old. In comparison with MT efforts in other countries, it would seem that Indian MT has a long way to go. However, this can also be very advantageous, because Indian researchers can learn from the experience of their global counterparts. In this paper, we discussed and looked at the various Machine translation systems in India along with their features and domain areas/applications in details. We also discussed the various approaches that are applied for building machine translation systems. Many researchers and research group have come up with different translation systems applying different approaches. It is concluded that direct approach for Machine Translation is most suitable for closely related languages i.e. the languages with similar structure. The transfer approach and statistical approach is suitable for languages with different structures. The best approach is the hybrid approach with highest accuracy which is a combination of rule based and statistical approach.

REFERENCES

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

[2] Sudip Naskar & Shivaji Bandyopadhyay, (2005) "Use of Machine Translation in India: Current status" AAMT Journal, pp. 25-31.

[3] Smriti Singh, Mrugank Dalal, Vishal Vachhani, Pushpak Bhattacharyya, Om P. Damani "Hindi Generation from Interlingua (UNL)" Indian Institute of Technology, Bombay (India)

[4] Ananthakrishnan R, Kavitha M, Jayprasad J Hegde, Chandra Shekhar, Ritesh Shah, Sawani Bade & Sasikumar M, (2006) "MaTra: A Practical Approach to Fully-Automatic Indicative English-Hindi Machine Translation", *in proceedings of the first national symposium on Modelling and shallow parsing of Indian languages (MSPIL-06)*organized by IIT BOmbay, 202.141.152.9/clir/papers/matra_mspil06.pdf

[5] Ruchika A. Sinhal, Kapil O. Gupta (2014) "A Pure EBMT Approach for English to Hindi Sentence Translation System" I.J.Modern Education and Computer Science, 2014, 7, 1-8 Published Online July 2014 in MECS (http://www.mecs-press.org/)

[6] G. S. Josan & G. S. Lehal, (2008) "A Punjabi to Hindi Machine Translation System", *in proceedings of COLING-2008*: Companion volume: Posters and Demonstrations, Manchester,UK, pp. 157-160.

[7] Vishal Goyal & Gurpreet Singh Lehal, (2010) "Web Based Hindi to Punjabi Machine Translation System", *InternationalJournal of Emerging Technologies in Web Intelligence*, Vol. 2, no. 2, pp.148-151, ACADEMY PUBLISHER

[8] Vishal Goyal & Gurpreet Singh Lehal, (2011) "Hindi to Punjabi Machine Translation System", *in proceedings of the ACL-HLT 2011 System Demonstrations*, pages 1–6, Portland, Oregon, USA, 21 June 2011

[9] Murthy. K, (2002) "MAT: A Machine Assisted Translation System", *In Proceedings of Symposium on Translation Support System(STRANS-2002),* IIT Kanpur. pp. 134-139.

[10] R.M.K. Sinha & A. Jain, (2002) "AnglaHindi: An English to Hindi Machine-Aided Translation System", *International Conference AMTA(Association of Machine Translation in the Americas)*

[11] Vishal Goyal & Gurpreet Singh Lehal, (2009) "Advances in Machine Translation Systems",National Open Access Journal, Volume 9, ISSN 1930-2940 http://www.languageinindia

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

[13] 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.

[14] Yanjun Ma, John Tinsley, Hany Hassan, Jinhua Du & Andy Way, (2008) "Exploiting Alignment Techniques in MATREX: the DCU Machine Translation System for IWSLT 2008', *in proceedings of IWSLT 2008*, Hawaii, USA

[15] Sanjay Chatterji, Anupam Basu (2012), "Repairing Bengali Verb Chunks for Improved Bengali to Hindi Machine Translation"

[16] M.L.Dhore & S.X.Dixit (2011) "English to Devnagari Translation for UI Labels of Commercial web based Interactive Applications"

[17] Devika Pishartoy, Priya, Sayli Wandkar (2012) "Exteneding capabilities of English to Marathi machine Translator"

[18] Charugatra Tidke, Shital B, Shivani P (2013) "Inflection Rules for English to Marathi Machine Translation"

[19] Abhay A, Anita G, Paurnima T, Prajakta G (2013), "Rule based English to Marathi translation of Assertive sentence"

[20] Krushnadeo B, Vinod W, S.V.Phulari, B.S.Kankate (2014), "A novel approach for Interlingual example-based translation of English to Marathi"

[21] G.V.Gajre, G.Kharate, H. Kulkarni (2014), "Transmuter: An approach to Rule-based English to Marathi Machine Translation" (2014)

[22] http://www.cdac.in/html/aai/mantra.asp

[23] http://ebmt.serc.iisc.ernet.in/mt/login.html

[24] http://shakti.iiit.net

[25] http://cdac.in/index.aspx?id=mc_mat_machine_aided_translation

[26] http://www.academia.edu/7986160/Machine_Translation_of_Bilingual_Hindi-English_ Hinglish_Text

[27] http://www.academia.edu/3275565/Lattice_Based_Lexical_Transfer_in_Bengali_Hindi_MachineS_ Translation Framework

