# Study and Review of Various Machine Translation Techniques

Neha A. Chavan*, Pravin K. Patil

*Department of Computer Engineering, Shram Sadhana Bombay Trust's College of Engineering and Technology, Bhambhori, Jalgaon-425001, Maharashtra*

*padmakargchavan@gmail.com*

***Abstract:***

*Machine translation(MT) can be described as the task of translating text or speech from one language to another with a little human effort as possible. MT aims to achieve quality translations which are semantically equivalent to source sentence and syntactically correct in the target language. MT simply performs the substitutions of words on the ground level. As the structure of every language is different there is huge gap between people of various cultures and society. The importance of MT is arising from the sociopolitical importance of translation in communities where more than one language is spoken. Every language has some rich literary text associated with it. But as it is in a regional language, so it cannot be understood by everyone. People those have the awareness of the said language only that will understand. MT breaks that barrier. The present paper deals with the sound collection of MT approaches in detail with sequential updates.*

***Keywords:*** *Machine translation, Rule based system, Corpus based system, Natural Language processing*

*Corresponding Author: padmakargchavan@gmail.com

## 1. Introduction:

Communication gap between people of various cultures and societies may arises because of Language only. In the field of Computer Science/Engineering Natural Language Processing (NLP) is the field that can fill this gap [1,2]. In the year 1965, Georgetown University and IBM jointly demonstrated Machine Translation (MT) system with their first successful attempt. Sometimes, culture may undergo into the process of vanishing which leads due to the loss/unrecognition of a language. Hence, language translation has much more importance. By the nature of human being one can most fluently express this thoughts/ideas/innovation in his native language. The best creations to the date by several authors and poets are in their regional languages for example Ravindranath Tagore. He used the Bengali language for his poems and songs. However, his poems and songs need to be converted to Hindi or English so that it can easily understood by everyone [3, 4].

Now a day, English language is used in most of the information communication systems. In fact, very few percentage of people can understand English language very easily. For the benefit of the society it is quite important to translate it into the several languages, particularly, National language of respective country. As like the example of Ravindranath Tagore, new theorem in a particular field has been discovered by Chinese scientist. This invention should be shared with all people in the world. However, due the language barrier it is not possible to understand. If this theorem has converted into the English language by means of MT it will be more beneficial from the point of view of social welfare [2, 5, 6].Bi-lingual speaker is required when there is a language barrier between two individuals and which is seen to be time consuming and costly too. However, MT can offer you with best solution for language barrier which will be cost effective and swift [1,7]. In India, Hindi language (principal official language of India) is widely used in addition to the English language. Thus, it is necessity to design a translator for converting one to other (Particularly, English to Hindi). At present, the Government of India is also made the efforts on the awareness of use of Hindi language.

In the above context, it has been cleared that the MT is quite important for the scientific and technological development/enhancement. The present paper is the collection of new insights of MT which is considered to be the further challenge to scientist of Computer Engineering.

## 2. Brief history of MT:

The information summarizes here is the historical background of MT [8-12]. The first idea was proposed by Weaver in 1949 regarding the translation by using computers. IBM in collaboration with research students from Georgetown University developed first basic automatic Russian-English translator in year 1544. Formulation of Automatic Language Processing Advisory Committee (ALPAC) with American government is done in year 1964 to study the perspectives and possibilities of MT. A well know Japanese firm 'SHARP' developed Automatic translator for English to Japanese language as DUET, which was based on the approach of rules and transfer. In year 2005, Google launched the first web site for automatic translation.

## 3. MT APPROACHES:

Existing MT systems along with various approaches are summarizes in Figure 1.

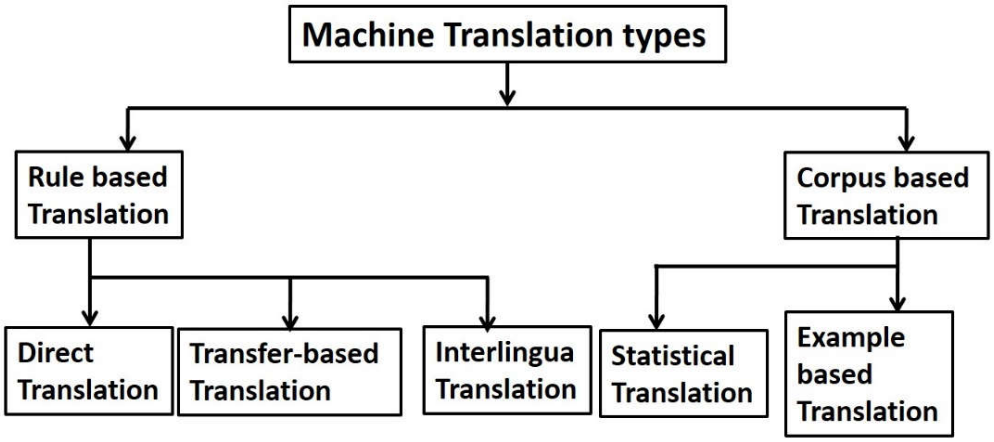

**Figure 1**. Machine Translation systems.

### 3.1 Rule based translation:

Rule based translation is working on predefined rules. The origin of these rules are grammar and dictionary of source and target language. Further, the source text is parsed and intermediate representation is produced. Finally, after applying rules to intermediated representation the target language is generated.

### 3.1.1 Direct MT Systems:

Direct word by word translation of source language is done in direct machine translation. The targeted code is produced after syntactical rearrangement.

**Example: Hindi-to-Punjabi MT System**

At Punjabi University, Patiala the translation system has been developed and the accuracy of the system based on intelligibility test is found to be 95.40%. Also, based on accuracy test it has been found to be 87.60%. [13, 14].

### 3.1.2 Transfer-Based MT Systems:

Various transfer rules are considered in case of transfer-based approach. For example, direct conversion into target language is takes place only when a sentence matches one of the transfer rules.

Example: Sampark System

A huge expectation based on Transfer-Based machine system has been completed by association of 11 Indian organizations. Their object was to establish multipart MT method for one language to other language used in India. For analyzing language and combining it with machine learning it uses Computational Paninian Grammar (CPG). Efforts of consortium leads into the translation of 18 Indian language pairs [11].

### 3.1.3 Interlingua MT Systems:

Interlingua MT System is seems to be best choice for direct machine translation. Since the direct matching of word by word is carried out it is observed to be less efficient. Intermediate structure is generated in this approach and is called as Interlingua which further gives more than one target language.

**Example: Angla Hindi (2003)**

For English to Indian languages, Angla Hindi MT System has been developed by Sinha et. al. All the modules of Angla Bharti and abstracted example-base translation has been used. The accuracy of translation for this system is 90% [15].

### 3.2 Corpus based MT system:

Corpus is nothing, but it is the collection of written test. For initial analysis a huge data is required. Corpus-based MT has two types as given below.

### 3.2.1 Example Based MT (EBMT) Systems:

**Example: The MATREX System**

DCU MT System for ICON 2008 was developed by Kumar et.al. Marker-based chunking is used in the MATREX system. At each new occurrence of a marker word the chunk is created in terms that at least one content word has to be shown by each chunk. The "edit-distance style" has been used by this system for the alignment of chucks [16, 17].

### 3.2.2 Statistical MT Systems:

**Example: English to Indian Languages MT System**

A consortium of nine institutions has developed the English to Indian Languages in Tourism and Healthcare Domains (EILMT) which is the MT System. For the baseline system the engine was initially developed. For testing and tuning training corpus consisted of 5000 sentences and 800 sentences were divided. It is observed that in producing a good quality translation deficient baseline technique was used in this system. Therefore, inclusion of pre-processing stage was done. The role of it is to look out the syntactic re-ordering on the source language. In addition, to separate out the root word and the affixes rule-based suffix separation approach was used. For English-Marathi and English-Bengali pairs the system is extended and tested [18].

## 4. Comparison OF MT approaches:

The following table shows the comparison between rule-based and corpus-based systems.

Table 1. Comparison of major MT approaches.

| MT Approaches | Merits | Demerits |
|---|---|---|
| **Rule-based** | • initial system can build Easily<br>• use of linguistic theories<br>• efficient for fundamental phenomena<br>• appropriate for domain specific translation<br>• superior translation quality | • experts are assigned to articulated rules<br>• Hard to maintain and extend<br>• Unproductive for managerial phenomena<br>• In case of general translation systems number of rules will raise in huge extend |
| **Corpus-based** | • Source of knowledge is corpus<br>• Built on translation patterns<br>• cost effective<br>• mathematically formulated | • linguistic background is absent<br>• high search cost<br>• existence of Information gaining problem<br>• parallel corpora are mandatory in large form<br>• irregular translation quality |

## 5. Conclusions:

In the present paper, the emphasis is given on MT for languages in India and other than India. Significant conclusions are summarized as given below,

➢ For purpose of tourism and health care Indian and non-Indian MT systems were established. More emphasis is made on better performance and accuracy as is given by rule-based method.

➢ For the Indian languages which are morphologically rich in features and agglutinative in nature Rule-based method is found to be unsuccessful in conditions where completeness of general-purpose MT systems is considered.

➢ It is also notice that for developing rule-based MT systems support of linguistic experts is essential.

➢ For many Indian languages even, monolingual corpus in not available whereas, enormous bilingual parallel corpora are necessary for Example-based as well as Statistical MT methods.

➢ It is clearly seen that not so far development is done regarding English to Marathi translation system which further, opens the door for future work.

## References:

[1] N. Sharma, "English to Hindi Statistical MT System" http://hdl.handle.net/10266/1449.

[2] R. Gupta, N. Joshi and I. Mathur, "Analysing Quality of English-Hindi MT Engine Outputs Using Bayesian Classification", International Journal of Artificial Intelligence & Applications, vol. 4, no. 4, (2013), pp. 165-171.

[3] N. Tomer, D. Sinha, P. Kant Rai, “F-Measure Metricfor English to Hindi Language MT”, vol. 1, no. 7, (2012), 151-156.

[4] O. Bojar, P. Stranák, D. Zeman, “Data Issues in Englishto-Hindi MT”, Proceedings of the LREC, (2010), 1771-1777.

[5] G. K. Anumanchipalli, L. c. Oliveira, A. w black, “Intent transfer in speech-to-speech MT”, IEEE Spoken Language Technology Workshop (SLT), Miami, FL, (2012) December, pp. 153-158.

[6] U. Muegge, “How to implement MT”, Lebende Sprachen vol. 3, (2202), pp 110-114

[7] E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Déchelotte, M. Federico, M. Kolss, Y.-S. Lee, J. B. Mariño, M. Paulik, S. Roukos, H. Schwenk, and H. Ney , “System Combination for MT of Spoken and Written Language”, IEEE, vol.16, no. 7, (2008).

[8] J. Hutchins, “Current commercial MT systems and computer-based translation tools: system types and their uses”, International Journal of Translation vol.17, no.1-2, (2005), pp.5-38.

[9] J. Hutchins, “The history of machine translation in a nutshell”, hutchinsweb.me.uk/Nutshell-2005.pdf

[10] J. Hutchins, “Historical survey of MT in Eastern and Central Europe”, Based on an unpublished presentation at the conference on Cross lingual Language Technology in service of an integrated multilingual Europe, Hamburg, Germany, (2012) May pp. 4-5.

[11] Sampark: MT System among Indian languages http://tdildc.in/index.php?option=com_vertical&parentid=74, http://sampark.iiit.ac.in/. (2009).

[12] A. Bharti, C. Vineet, A. P. Kulkarni & R. Sangal,” ANUSAARAKA: overcoming the language barrier in India”, published in Anuvad: approaches to Translation, (2001), pp. 1-19.

[13] Sitender & S. Bawa, “Survey of Indian MT Systems”, International Journal Computer Science and Technolgy, vol. 3, no. 1, (2012), pp. 286-290.

[14] G. S. Josan & J. Kaur, “Punjabi To Hindi Statistical MachineTransliteration”, International Journal of Information Technology and Knowledge Management, vol. 4, no. 2, (2011), pp. 459-463.

[15] R. M. K. Sinha & A. Jain, “Angla Hindi: An English to Hindi Machine-Aided Translation System”, International Conference AMTA (Association of MT in the Americas), (2002), pp. 1-5.

[16] A. K. Srivastava, R. Haque, S. K. Naskar & A. Way, “The MATREX (MT using Example): The DCU MT System for ICON 2008”, in Proceedings of ICON-2008: 6th International Conference on Natural Language Processing, Macmillan Publishers, India, (2008), pp. 1-4.

[17] Y. Ma, J. Tinsley, H. Hassan, J. Du & A. Way, “Exploiting Alignment Techniques in MATREX: the DCU MT System for IWSLT 200”, in proceedings of IWSLT, Hawaii, USA, (2008), pp. 26-33.

[18] R. Ananthakrishnan, J. Hegde, P. Bhattacharyya, R. Shah & M. Sasikumar, “Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical MT”, in proceedings of International Joint Conference on NLP Hyderabad, India, (2008), pp. 513-520.