MachineTranslation and Approaches: Review

Dr. M. Narendra[1], Dr. K.M. Rayudu[2], Dr. T. Sivaratna Sai[3] Satyanarayana Vittal[4,] Narasimha Rao Nakka[5]

[1, 2, 3] Department of CSE, [4]Department of ECE, [5]Department of EEE, QIS College of Engineering and Technology, Ongole, AP, India, qispublications@qiscet.edu.in

Article Info
Page Number: 161-175
Publication Issue:
Vol 69 No. 1 (2020)

ABSTRACT:
Code-mixing is the blending of at least two dialects or language assortments in discourse. It is utilized to allude to expressions that draw from components of at least two syntactic frameworks. Some work characterizes code-blending as the setting or blending of different phonetic units (attaches, words, phrases, provisos) from two distinct linguistic frameworks inside a similar sentence and discourse setting. This arising method of correspondence is turning out to be essential for everyday life correspondence in multilingual nations like India. Investigating these kinds of correspondences are very simple for people however for machines it turns into an intricate errand. The outcome of this paper is to examine various methodologies of machine translation, their pertinence for code-mixing and furthermore the moves that should be tended to. It likewise presents rules for future examination work.

KEYWORDS:
Machine translation (MT), Code-mixing, Language Analysis, Hinglish, Corpus BasedMT, and Rule based MT, Hybrid MT

Article Received: 10 August 2020
Revised: 15 September 2020
Accepted: 20 October 2020
Publication: 25 November 2020

1. INTRODUCTION

In a multicultural nation like India which has approximately 840+ languages, communication using pure language is decreasing day by day in this digital world. People nowadays are using a mixture of more than one language in order to express their thoughts, feelings, emotions, etc. Such a combination introduces mixed languages such as Hinglish (mixture of Hindi and English) and many more. Code-mixing of Hindi and English is becoming a common phenomenon nowadays, due to which people have started using it as a separate language.

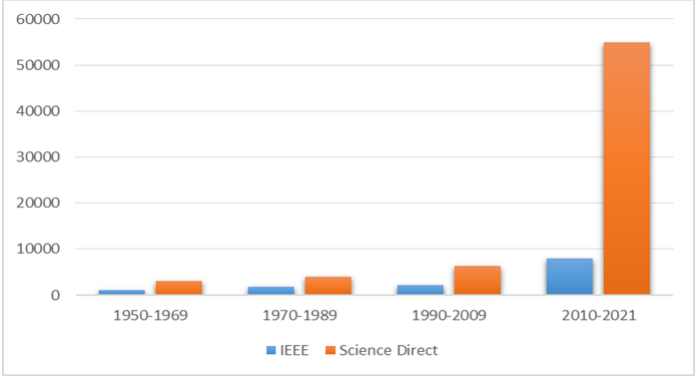


Fig 1.1 Diagram of work finished till date on machine translation on reputed paper

publishing sites like IEEE, Science Direct.

Some example given below show the degree of mixing.

(1) Usane dirty cloths wash kiye

(2) Usane newspaperoN KO bahut carefully read kiyaa.

(3) Usane bahut practice ki, but did not understand the Raagas.

(4) कल afternoon 3 pm मिलते है

From above examples it can be clearly seen that code mixing in Hindi-English is diverse. Thesituation is becoming so common that people are considering it a distinct language and named it Hinglish. And thus analysis of such content becomes important. Even though there is a lot of work done on Machine Translation already, it is mostly restricted to pure languages only. These approaches can't have the same efficiency when applied to code-mixed languages as they have on pure languages. Also, with thorough study and facts, it can be clearly seen that there is an increase in research work and also the need for work on this topic has increased in these years. Research and study on this topic will surely make a huge difference when it comes to Data Analytics particularly for multilingual countries. It seems too easy with human intervention; but in the era of automation, AI and considering generation of data every day in the current scenario, it is needed to have automated versions of such systems.

2. VARIOUS APPROACHES USING MACHINE TRANSLATION

A machine translation framework will initially examine the source language input and an insideportrayal of it will be made. The portrayal will then, at that point, be controlled and moved tosuit the objective language. Output should be Target Language.

Two approaches of machine translation are as follows:

1) Rule based MT(RBMT)

2) Corpus based MT

In RBMT explicit arrangement of rules are characterized by human specialists for interpretation and for this enormous measure of info is needed to be given. However in corpus-based methodology there is the programmed extraction of information by dissecting interpretation models from an equal corpus worked by Human experts. Hybrid MT method is nothing but a combination of elements of two significant sorts of MT.

MACHINE TRANSLATION (MT)

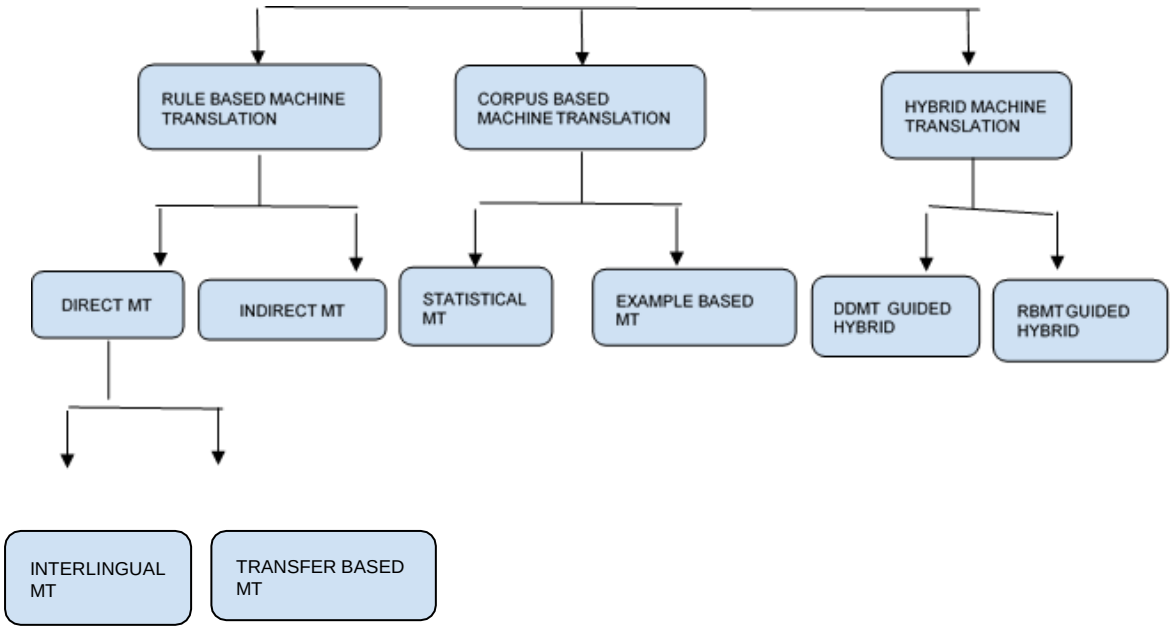


Fig 2.1: Approaches in Machine Translation

2.1. Rule based Approach

RBMT is also known as Knowledge-Based Machine Translation is an exemplary methodology wherein interpretation depends on etymological data about source language and target language that is for the most part taken from dictionaries(bilingual) and depend on punctuations covering the principles for morphology, linguistic structure, lexical and semantic analysis and so known as rule based.

Morphology analysis- It is a branch of linguistics that involves the study of words, their

formation, and their relationship to other words in the same language.

Syntax Analysis: It checks whether a sentence is well-formed according to the rules and structure of formal grammar. For example, 'goes school the boy' will be rejected as the sentenceis grammatically incorrect.

Semantic Analysis: Meaningfulness of the text will be checked. Exact/Dictionary meaning isdrawn from the text in this phase. Example: 'Boy ate tea' will be rejected.

Lexical selection: In this phase, we access and fit the appropriate words in a given text.

2.1.1. Basic Principles of RBMT Aprroach

The arrangement of etymological standards in RBMT are applied in three levels:

1. Analysis
2. Transfer
3. Generation

Thus it requires punctuation examination, semantic investigation, sentence structure formation,and semantic generation. RBMT produces the objective message by following the means shown:

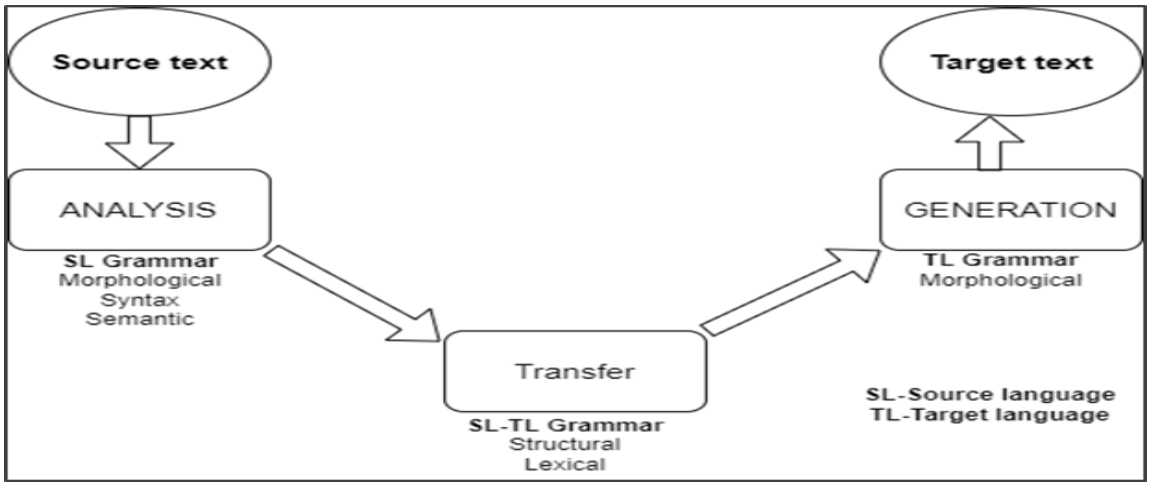


Fig 2.2. Architecture of RBMT

The fundamental technique of working in RBMT is connecting the layout in a given sentenceto the construction of the normal end result sentence without affecting its meaning.

The accompanying model can exhibit the working of RBMT Source Language = English; Demanded Target Language = Hindi

1. A word reference that will plan every English word to a fitting Hindi word.
2. The Rules addressing customary structure of English sentence.
3. The Rules addressing customary structure of Hindi sentence.

At last the rules which can correlate both construction of sentences are needed to give finaloutput.

2.1.2. Challenges in RBMT approach

- Limited dictionaries are available and building a new dictionary is costly.
- Some linguistic data have to be set manually
- It becomes complex to interpret idioms using RBMT

2.2. Sub Approaches under RBMT

2.2.1 Direct machine translation:

It is one way bilingual MT which requires least structure examination of source language. [2, 3, 4, 5] It is likewise called dictionary-based translation. Direct translation is achieved by the usage of 4 steps:

1. Identification of base form of the word in the SL and resolve ambiguity using Morphological analysis.
2. Bilingual word references are alluded to interpreting the words in the SL to create comparable words of objective language
3. Minor grammatical adjustment of Target language and TL morphology generator is done using certain rules in grammar.
4. Output is a TL text

Example: I ate food. (SL-English)

ि ै ं खाना खाया. (Word-to-Word translation)

ि ै ं ने खाना खाया. (TL-Hindi)

morphological analyzer, reordering rules and bilingual word reference determines the overall quality and quantity of DBMT.

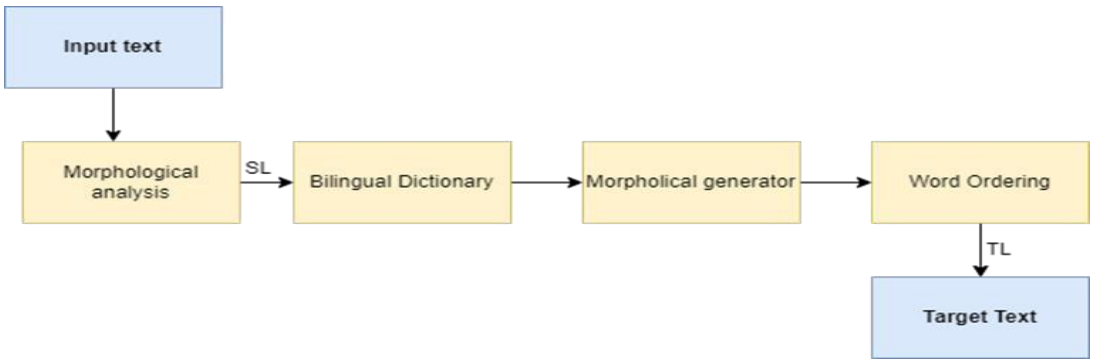


Fig 2.3. Architecture of Direct MT

2.2.1.1. Challenges in Direct Machine translation

1. It is word to word translation with least order adjustment and final output may havesome linguistic errors, improper order, partial understanding of SL.

2. Missing linguistic analysis between principle parts of sentences.

2.2.2. Indirect Translation

In Indirect translation basic assessment which joins Morphology, semantics and syntactic examination are applied over every SL message and changed over to momentary depiction which is dominatingly in kind of theoretical parse tree, and thereafter target message is accomplished through essential change reliant upon the specific generator as displayed in the Vaquois triangle in fig 2.4.

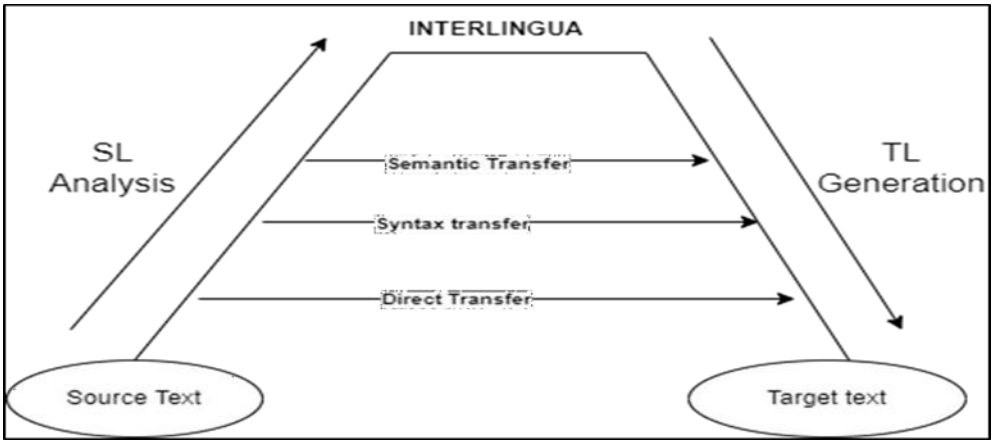


Fig 2.4. Vaquois triangle

2.2.2.1 Interlingual Approach

In the Interlingual machine interpretation approach, the SL is converted to a mediator language known as Interlingua. Interlingua is outlined by the mix of two words inter which suggests mediator and lingua infers language portrayed as a homogenous, hypothetical, unambiguous, and independent language [6, 10, 11, 12]. The TL is then generated from interlingua. The quantity of target language it tends to be changed into increments because of which interlingua turns out to be more important and is a significant benefit of this methodology.[5, 8, 9].

Example: To Translate from Japanese to Hindi; the English language is used as an InterlinguaOnamae wa nan desu ka(SL – Japanese)

What is your name? (interlingua-English)

तुम्हारा नाि क्या हैं?(TL-Hindi)

Challenges in Interlingual translation approach:

1. Defining Interlingua becomes tough even for languages which are similar. (E.g. Romanlanguages)

2. Extracting meaning from Source language is complex to generate Interlingua.

2.2.2.2. Transfer based Machine learning approach

In order to overcome disadvantages in the Interlingual approach transfer approach was invented. It also creates translation from intermediate representation. Unlike the Interlingual approach, it has two portrayals one identified with SL and one more identified with TL comprises three principle stages that are Analysis, synthesis, transfer.

The analysis is done using linguistic information and various algorithms are applied on SL to get syntactic and semantic structure. Syntactic portrayal of a SL sentence is created utilizing a SL parser.

In the transfer phase, the structure of the SL is converted to equivalent TL structure using some transfer rules which are also called oriented representations [4, 6, 8, 10, 11] .

For example:

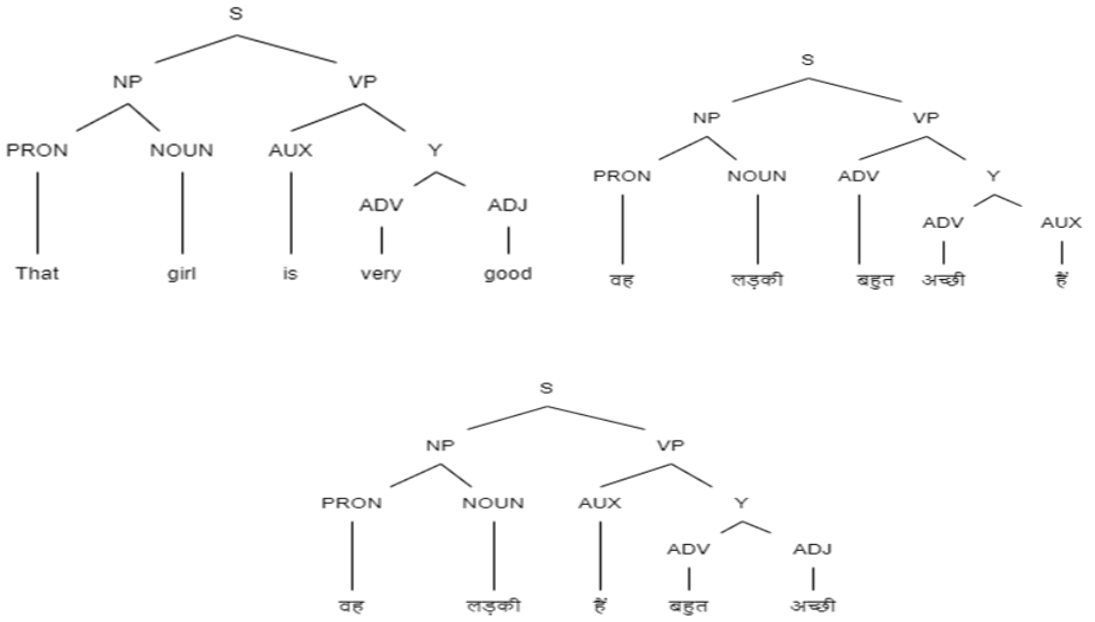


Fig 2.5 Structural Transfer in Hindi-English Machine Translation System

TL is generated in last phase with the help of TL structure and using morphological analyzer final TL sentences are generated. Fairly high-quality translations are obtained using this approach, with an accuracy of around 90%.

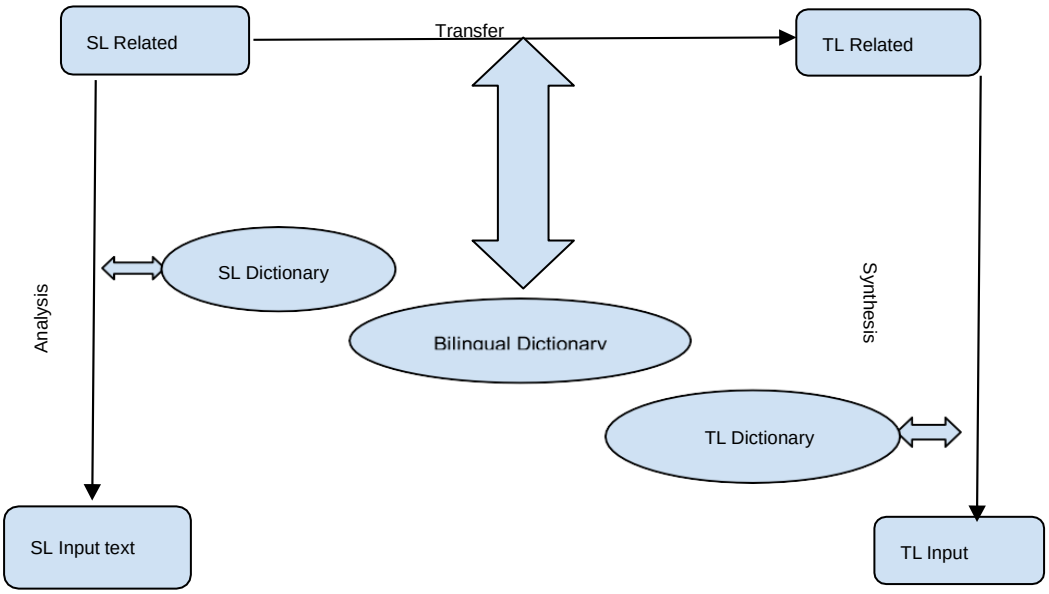


Fig 6. Transfer based approach

Challenges in transfer-based approach:

1. Rules must be applied at every stage of interpretation. There are rules for analysis of SL,TL and also for source to target transfer.

2. It is hard to keep transfer modules as straightforward as could really be expected.

3. CORPUS BASED MACHINE TRANSLATION

3.1. Statistical machine translation (SMT)

SMT is an information based approach and it uses parallel aligned corpora. A translation in every sentence is a target language for causing a Mathematical reasoning problem. A probability and accuracy of translation are directly proportional to each other

The SMT architecture of SMT comprises of 3 models [2, 6, 13, 14]:

- P (t) which is used to calculate the TL probability.
- P (t|s) is conditional probability of TL if SL is given as input.
- Decoder model is used to get best translation if two probabilities given above are maximized also given in below equation and also makes use of search algorithms.

t= arg max (p(t|s)*p(t)

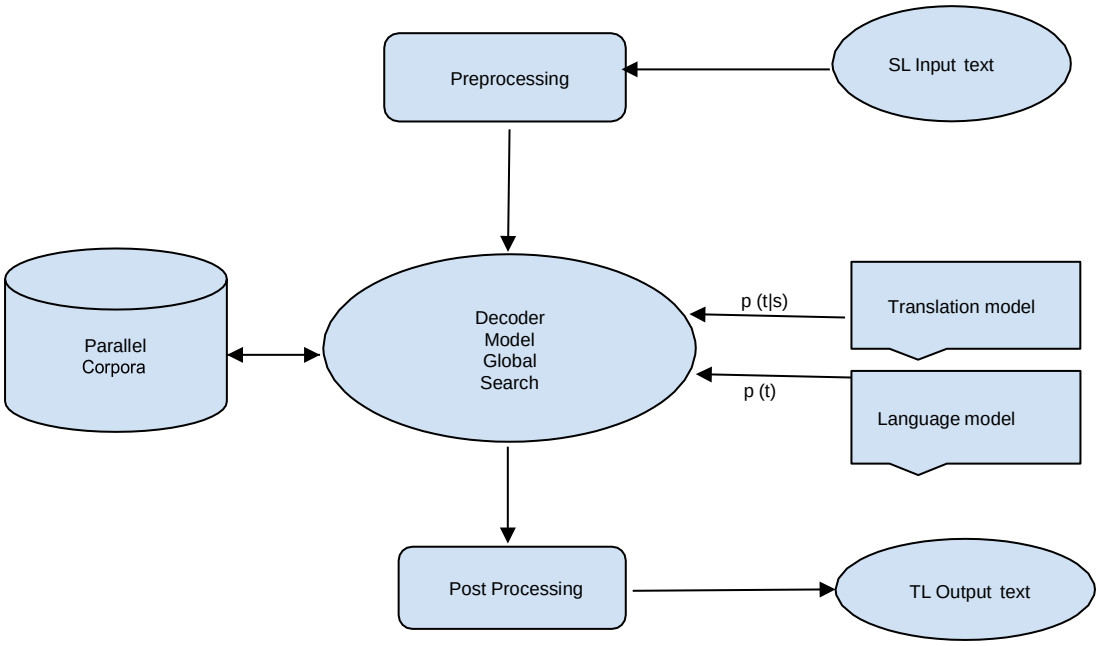


Fig 7: SMT Architecture

There are three approaches:

(1)Word based (2)Phrase based (3)Hierarchical based

1. Word-based SMT:

Every word in SL sentence is translation to TL. The translated words are then rearranged inmeaningful order using some reordering algorithm for generation of final output in TL

2. Phrase-based SMT:

Phrases as the major unit of interpretation were proposed by Koehn [15]. By interpreting entire arrangements of words, the length might contrast to reduce the limitations of word-based interpretation. This sort of translation is called Block. The phrase-based translation models are derived from the parallel corpus by removing all similar solid expressions with this term basedon Koehn's principle [15]. As Antony recommended, the input and output clauses are adjustedaccordingly.

3. Hierarchical phrase-based SMT: It is a combination of word based and phrase based model. This model was built by Chiang. Phrase based include block for translation and wordbased involves rules.

Challenges of Statistical Machine Translation Approach

- ➢ Restricted assets are costly for Corpus Creation.
- ➢ The outcomes are unforeseen. Shallow familiarity can be misleading.
- ➢ Between dialects that have fundamentally unique word orders, Statistical Machine Translation doesn't function admirably.
- ➢ The advantages are overemphasized for European dialects.

3.2 Example-based Machine Translation:

Primary information on EBMT is portrayed by the utilization of bilingual corpus with parallel texts. It tends to be considered to be an implementation of the case-based reasoning approach. An EBMT framework is given a bunch of sentences in the SL and corresponding interpretations of each sentence in the TL with highlight point planning. EBMT is utilized to make an interpretation of SL to TL by the utilization of comparative kinds of sentences.

There are four tasks in EBMT:

- example acquisition
- example base and management
- example application
- Synthesis.

EBMT is the interpretation by analogy. The rule of interpretation by analogy is encoded to EBMT through the example translations that are utilized to prepare such a framework.

Example:

English	Hindi
“Rama sings a song”	“Rama geet gaati hai”
“Ronit sings a song”	“Ronit geet gaata hai”

In this example Rohit is replaced with Rama and Gata is replaced with gaati and then the finaltranslation is done.

When we have to translate structurally similar sentence of SL into a different structure then thistype of translation may not be effective

Example:

English	Hindi
How much is that red umbrella?	wah laal rang ka chata kitne ka hain?
How much is that small camera?	Wah chota kaimara kitne ka hai?

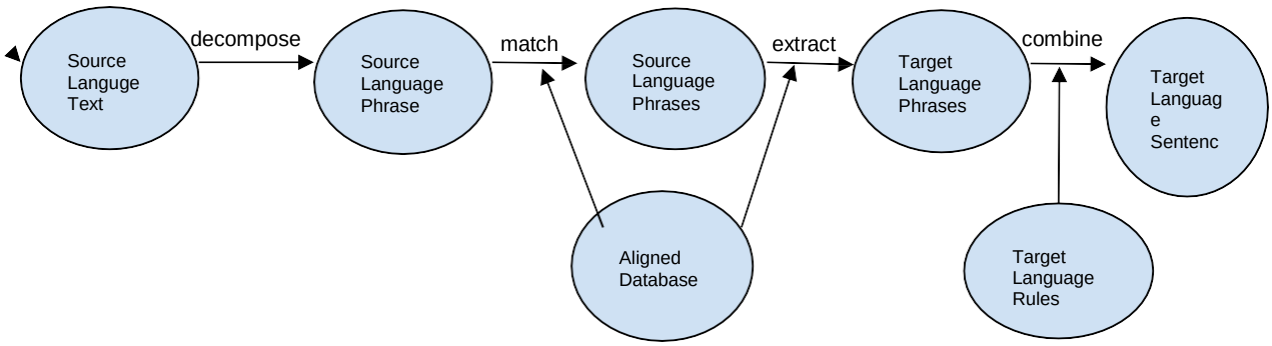
The sentence pairs as illustrated in an example shown in table depicts the way in which EBMTare trained from bilingual parallel corpora

In table (2) example containing sentences contain in one language translated into another language. It is example of minimal pair which means only one element is the sentences different.

For example, an EBMT system would learn these units of translation from the above example[16]:

In table (2) How much is that X? Corresponds to wah X kitne ka hai. Red umbrella corresponds to laal chata.

Small camera corresponds to chota kaimara.



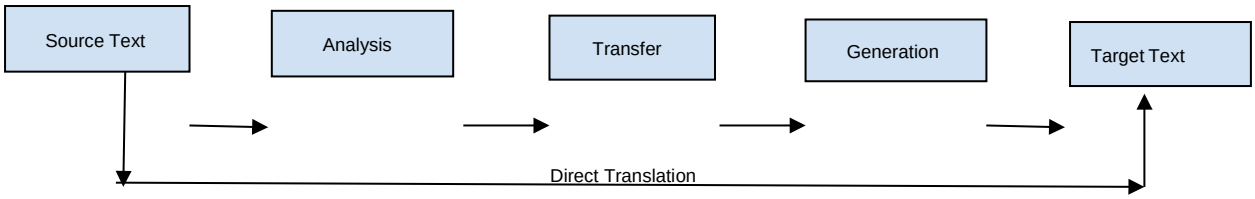


Fig 3.1:- Example based Machine Translation

Fig 3.2:- Example Based MT Flow

The Example-based translation uses three stages;

1. Analysis or Matching
2. Adaption or Transfer
3. Recombination or Generation

1. Matching or Analysis:

Source Input text is categorized according to the granularity and is followed by search for modelsfrom the knowledge base that are or are closely related to the character unit of the source language category and the key character unit is selected. The objective language sections relatingto the significant pieces are separated.

2. Adaption or Transfer:

If the analysis component is accurate then the fragments are reassembled to form a TL output orit will receive a TL component of the appropriate analysis corresponding to a particular component in SL and aligned.

3. Recombination or Generation:

It is a combination of TL fragments suitable for creating the official target text.

Challenges of EBMT approach:

- It is a dynamic way to approach translation because it discards the need for manuallyderived rules.
- To generate the dependency trees EBMT requires analysis and generation modules.
- Parallel computation techniques can be applied to EBMT but another problem with EBMT is computational efficiency, especially for large databases.

4. HYBRID MACHINE TRANSLATION APPROACH

Hybrid approach is combination of RBMT as well as CBMT approach which has more accuracy in translation. As of now, this approach is used in government as well as private sector related to machine translation.

We can utilize this approach in various way. Sometimes we can use rules to process the data

firstand then correct the final output with the help of statistical calculations. In other cases rules are also used to process the output which comes from statistical phase. The later is more powerful than former.

Two methodologies are conceivable in Hybrid MT

4.1 RBMT guided Hybrid

The methodologies incorporates the utilization of expressions and model taken from equal corpusto upgrade the word reference.

4.2 DDMT guided hybrid

Rules are used either in pre handling/post handling or at the Centre of the model.

Sr. No	Approach	Accuracy Range
1	Rule based Machine translation	60 % to 77.7%
2	Corpus Machine translated	60% to 80%
3	Hybrid Machine translation	80% to 90%

Table 4.2.1. Accuracy of Various Approaches of Machine Translation

5. Future Research Directions

- Direct Machine approach translates word to word from SL to TL with basic analysis and less consideration of grammar. It works well only for small sentences. We could work to improve its performance to translate group of sentences or paragraph along with more grammatical aspects so that it will produce better results.
- In the interlingual language translation approach, an intermediary language is requiredwhich makes the process complex as the intermediate language generated may also have different grammar and this could bring ambiguity when interpreting the source language. We can work on it by eliminating intermediary language.
- All the approach mention above mostly work on pure language but for code mix languagewe need more systemic approach in machine translation also accuracy of the above mentioned approaches varies drastically even if slight change in grammar or spelling nuance results in low accuracy.
- Now a day's code mixing has become very common phenomenon. In the Era of globalization where people from vivid background interact, combination of two or more languages while communicating is becoming a normalcy. As a human we can interpret it but for machine understanding such languages is quite difficult so we can work on developing a system to analyse such combined languages e.g. Hinglish (English + Hindi).

6. Conclusion:

In this review paper importance, evolution and uses of machine translation is discussed. It also gives insights on various approaches of machine translation through rules of grammar, intermediate language or considering previous examples given to machine with their challenges. It also sums up previous research work that has been done on this topic and also provides future

research work that can be done. Hybrid approach in machine translation gives more accuracy than other two approaches. Nowadays, MT is focusing on translation of code mixed languages to pure language considering the former as standalone language. It is developing and challengingarea of NLP .We aim in development and expanding our study to handle translation of code mixing in the future.

7. References

1. IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 2, September2014 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784 www.IJCSI.org Indonesian Journal of Electrical Engineering and Computer Science 1(1):182
2. DOI:10.11591/ijeecs.v1.i1.pp182-190
3. Peng L. A Survey of Machine Translation Methods. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2013; 11(12): 7125-7130
4. Hutchins W.J, Somers H L. An introduction to machine translation. London: Academic Press.1992:
5. Slocum J. A survey of machine translation: its history, current status, and future prospects.Computational linguistics.1985;11(1):1-17
6. Antony P J. "Machine Translation Approaches and Survey for Indian Languages." International journal of Computational Linguistics and Chinese Language Processing. 2013; 18(1): 47-78
7. Peng L. A Survey of Machine Translation Methods. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2013; 11(12): 7125-7130
8. Chéragui, Mohamed Amine. Theoretical Overview of Machine Translation. Proceedings ICWIT.2012;
9. Hutchins John. A new era in machine translation research. In Aslib proceedings. 1995; 47(10)211-219
10. Tripath s , Sarkhel. K. Approaches to machine translations. Annals of Library and information
11. studies.2010; 57: 388-393.
12. Ansary S. Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas. In 11th

13. International Conference on Language Engineering, Ain Shams University, Cairo, Egypt. 2011:

14. Hiroshi U , Meiying Z. Interlingua for multilingual machine translation. Proceedings ofMT Summit

15. IV, Kobe, Japan. 1993:157-169.

16. Juss`a, M, Farru´s M, Marin´o.J, Fonollosa.J. study and comparison of rule- based and statistical Catalan-S panish MT systems Computing and Informatics. 2012; 31: 245–270

17. Saini Sandeep. Vineet Sahula. A Survey of Machine Translation Techniques and Systems for Indian Languages. In Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference.2015: 676-681.

18. Koehn P, Och J, Daniel Marcu. Statistical Phrase-Based Translation. Proceedings of HLT- NAACL,Edmonton, May-June 2003. Main Papers , 2003: 48-54.

19. MD Okpor. Machine translation approaches: issues and challenges. International Journal of Computer Science Issues (IJCSI), 11(5):159, 2014.

20. Béchara Hanna. Raphaël Rubino. Yifan He. Yanjun Ma. Josef van Genabith. An Evaluationof Statistical Post-Editing Systems Applied to RBMT and SMT Systems. In COLING. 2012; 21: 5-230.

21. Costa-Jussa Marta R. José AR Fonollosa. Latest trends in hybrid machine translation andits applications. Computer Speech & Language. 2015; 32(1): 3-10.

22. P Ramprakash, M Sakthivadivel, N Krishnaraj, J Ramprasath. ”Host-based Intrusion Detection System using Sequence of System Calls” International Journal of Engineering and Management Research, Vandana Publications, Volume 4, Issue 2, 241-247, 2014

23. N Krishnaraj, S Smys.”A multihoming ACO-MDV routing for maximum power efficiency in an IoT environment” Wireless Personal Communications 109 (1), 243-256, 2019.

24. N Krishnaraj, R Bhuvanesh Kumar, D Rajeshwar, T Sanjay Kumar, Implementation of energy aware modified distance vector routing protocol for energy efficiency in wireless sensor networks, 2020 International Conference on Inventive Computation Technologies (ICICT),201-204

25. Ibrahim, S. Jafar Ali, and M. Thangamani. "Enhanced singular value decomposition for prediction of drugs and diseases with hepatocellular carcinoma based on multi-source bat algorithm based random walk." Measurement 141 (2019): 176-183. https://doi.org/10.1016/j.measurement.2019.02.056

26. Ibrahim, Jafar Ali S., S. Rajasekar, Varsha, M. Karunakaran, K. Kasirajan, Kalyan NS Chakravarthy, V. Kumar, and K. J. Kaur. "Recent advances in performance and effect

of Zr doping with ZnO thin film sensor in ammonia vapour sensing." GLOBAL NEST JOURNAL 23, no. 4 (2021): 526-531. https://doi.org/10.30955/gnj.004020 , https://journal.gnest.org/publication/gnest_04020

27. N.S. Kalyan Chakravarthy, B. Karthikeyan, K. Alhaf Malik, D.Bujji Babbu,. K. Nithya S.Jafar Ali Ibrahim , Survey of Cooperative Routing Algorithms in Wireless Sensor Networks, Journal of Annals of the Romanian Society for Cell Biology ,5316-5320, 2021

28. Rajmohan, G, Chinnappan, CV, John William, AD, Chandrakrishan Balakrishnan, S, Anand Muthu, B, Manogaran, G. Revamping land coverage analysis using aerial satellite image mapping. Trans Emerging Tel Tech. 2021; 32:e3927. https://doi.org/10.1002/ett.3927

29. Vignesh, C.C., Sivaparthipan, C.B., Daniel, J.A. et al. Adjacent Node based Energetic Association Factor Routing Protocol in Wireless Sensor Networks. Wireless Pers Commun 119, 3255–3270 (2021). https://doi.org/10.1007/s11277-021-08397-0.

30. C Chandru Vignesh, S Karthik, Predicting the position of adjacent nodes with QoS in mobile ad hoc networks, Journal of Multimedia Tools and Applications, Springer US,Vol 79, 8445-8457,2020