Section A-Research paper

DIFFERENT WAYS TO ADDRESS MACHINE TRANSLATION USING MACHINE LEARNING FOR SANSKRIT LANGUAGE

ARCHANA SACHINDEO MAURYA
Institute of Technology (IoT), DOCSIS, SRMU, Lucknow.
*(Corresponding Author)
***PROMILA BAHADUR**
Associate Professor, Institute of Engineering & Technology (IET), Lucknow.
DIVAKAR YADAV
Professor, Institute of Engineering & Technology (IET), Lucknow

Article History: Received: 01.02.2023 **Revised:** 07.03.2023 **Accepted:** 10.04.2023

Abstract
One of the most significant uses of natural language processing is machine translation. It is a sophisticated computer-based translation technique that has been under investigation for many years. There has been a great deal of visible work done in this area. Before there were computers to apply to it, machine translation required significant assumptions. This study tries to present several machine translation methodologies, tools, and algorithms. The success rate and difficulties encountered are also discussed in the paper.

Keywords: Sanskrit, English-to-Sanskrit Machine Translation, Statistical Machine Translation, Rule-Based Machine Translation, Hybrid Approach

1. Introduction
Language is a common occurrence, a component that unites different communities, and a means by which people try to transmit their feelings and thoughts to others. There are some barriers or difficulties that translators encounter in this process, yet translation plays a crucial role in transferring societal concepts between at least two languages. We understand that translation plays a big part in removing impediments caused by various societies and correspondence. In this sense, translation is a fundamental, essential, and sufficient method of advancing society. In order to consider the sequential requests, communicate importance, and improve the relevant controls, records, and tight basis of the source text, a good translator must also be aware of social components, views, and norms. [1].

Translations into and out of other Indian languages use Sanskrit as an intermediary language. Sanskrit is the mother tongue of all Indian languages, it is true. In the Sanskrit language, machine translation has been tried in a variety of ways by linguistics. The approaches can be understood from a variety of perspectives, such as translation in terms of the technology, methods, and strategies employed [2].Additionally, the translated text is frequently estimated based on the factors listed below, which are detailed below.

Table 1: Criteria for evaluating the Translation

S. No.	Parameters	Description
1	Translation Speed	Based on how quickly the translated text responds.
2	Digital Content	Depending on how well the text, video and audio are translated across various devices.
3	Cross-Platform	Depending on how widely different platforms may embrace translation technologies.
4	Translation Quality	Based on the accuracy and quality of the translation
5	MT Approaches	Depending on how effective various methods are.
6	Cost	Depending on how much it ultimately cost to translate a document from one language to another.

2. Technology used for Machine Translation:

Figure 1 displays various kinds of classification for Machine Translation.

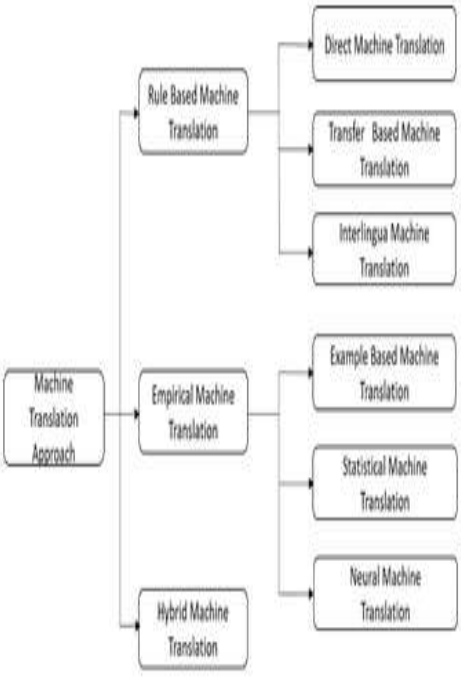

Figure 1: Classification of Machine Translations

Machine Translation can be classified into the following categories on the basis of used technologies:

- Machine Translation based on rules
- Machine Translation using evidence
- Machine Translation using hybridization

Figure 1 illustrates the further classifications of rule-based machine translation into Direct, Transfer-based, and Interlingua MT and of empirical machine translation into Example Based, Statistical Machine, and Neural Machine Translation.

2.1 Machine Translation Based on Rules

The size of a bilingual dictionary, the size of the rules, the parse tree, and the sentence analyzer are the primary components of rule-based machine translation (RBMT). Each language pair has multiple bilingual dictionaries and built-in linguistic rules. As seen in figure 2, the method has recorded a 90% accuracy rate [1].

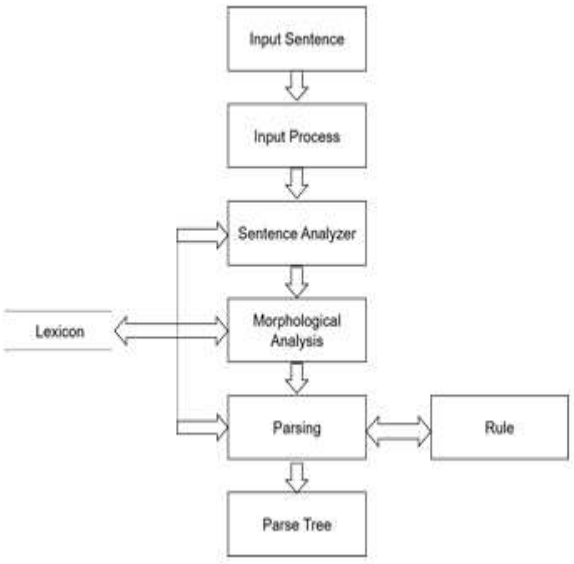

Figure 2: Machine Translation process based on rules

2.2 Machine Translation Based on Statistics

How big the training corpora are and what resources, such as dictionaries and other linguistic tools, can be used to assess the quality of statistical machine translation (SMT)? The language combination utilized affects the quality of the translation. Figure 3 illustrates the main components of SMT, including testing, decoding, training sets, and dictionaries. The technique has recorded an accuracy rate of 87% [3].

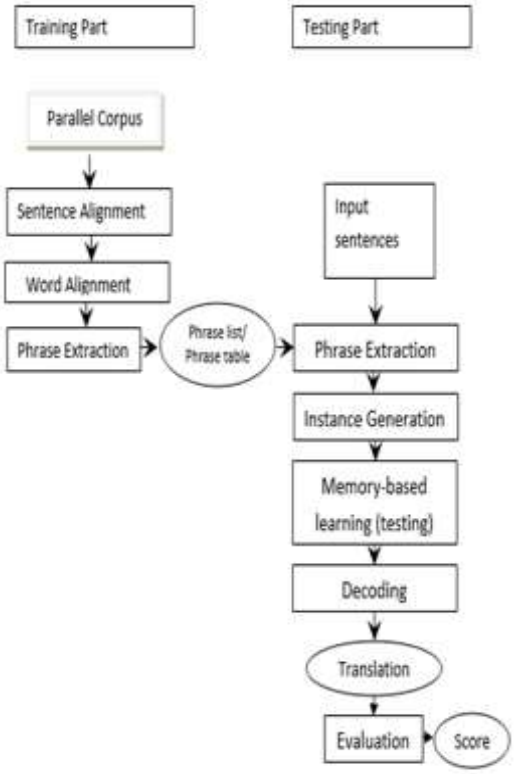

Figure 3: Machine Translation process using Statistics

2.3 Neural or Artificial Machine Translation

In Neural Machine Translation (NMT), a neural network model can be used to learn a statistical model for machine translation. In NMT, a single system can be trained directly on source as well as target, and it doesn't rely on any specific system. The model is shown in figure 4.

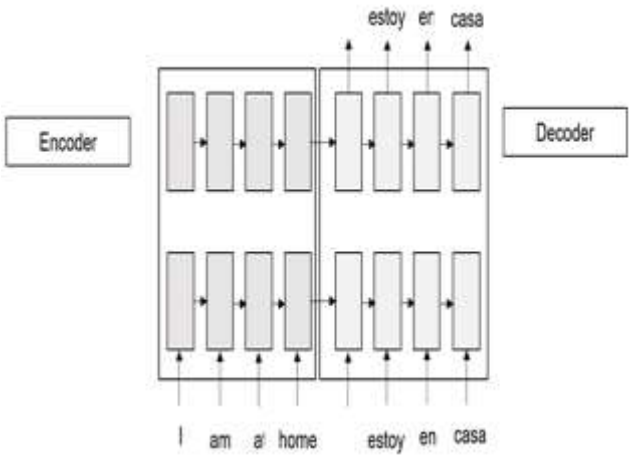

Figure 4: Machine Translation process by Artificial Neurons

2.4 Machine Translation using Hybridization

This is the combination of machine translation using rules and machine translation using statistics. This type of machine translation first integrates the data information into a rule-based translation, and later linguistic rules are integrated into a corpus-based architecture.

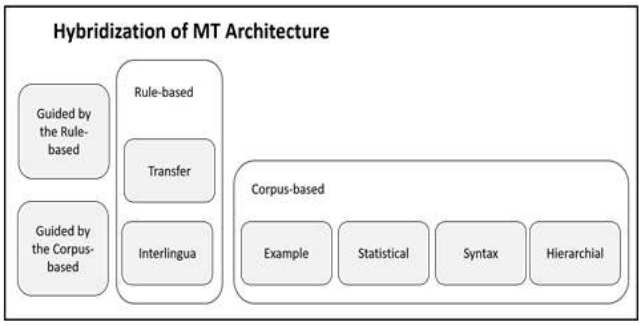

Figure 5: Hybrid Machine Translation process

3 Techniques Used For Machine Translation

a. In contrast to binary branching, phrase dependency tree-bank (PDT) includes flat structures. Instead of relying on syntactic functions, flat structure dependency is based on semantics [4].

b. The automatic suggestion of replacements of words for a machine translation output is proposed in the Automatic Post-editing Tool [5]. This method makes advantage of a multilingual word embedding technology.
Two lexical mistakes—"not translated word" and "incorrectly translated word"—are also used to illustrate how effective the tools are.

c. It is recommended to use constraint grammar (CG) rules in an Irish partial dependency parser. In an unconstrained Irish text, these rules are employed to provide grammatical functions and annotate dependency relations. The operation of chunks relies on dependency-tagged phrases and regular-expression grammar. Additionally, the system has stated that it achieves an f-score of 97.20 percent on development data and 93.50 percent on unseen test data, while the chucker earns an f-score of 93.60 percent on development data and 94.28 percent on test data that has not been seen.
A regular-expression grammar that works on the dependency-tagged phrases is used for chunking [6].

d. The enhancement of Sanskrit text to appearance in the Universal Networking Language (SANSUNL) was proposed. Improvements were made to the parsing, POS tagging, and Sanskrit language processing. Additionally, the suggested system stems the Sanskrit phrase using 774 suffixes and 23 prefixes with Sanskrit grammar rules. Efficiency was 95.375 percent [7] based on the BLEU and Fluency score measurements.

e. A virtual translation tool is provided that can produce text or speech in other languages. The system is meant to help viewers understand content in foreign languages during live presentations. In this method, neural machine translation replaces the traditional translator. Text to speech and speech recognition technologies have improved further human-machine interaction. The efficiency of the suggested system architecture and deployment strategy was demonstrated by the Vietnamese-English language pair [8].

f. The Dynamic Quality Framework (DQF) platform and ten expert translators were used to create the suggested system. The Microsoft Engine Translation (MET) and Google Neural Machine Translation provided raw machine translation parts of two texts, a marketing text and an instruction manual, to start (TAN). The productivity of the neural motor was also demonstrated by the results because post-editing suggestions required less time [9].

g. Reportedly, there are three translation ways that the Hinglish to Pure Hindi and English Translator tool can go: from Hinglish to pure Hindi and English, pure Hindi to pure English, and vice versa. The Direct MT Approach, Rule-Based MT Approach, and Hybrid MT Approach are the methods used for word order. Additionally, when the input sentences were in Hinglish, the computer recorded

accuracy rates of 91 percent for sentences provided in Hindi and 84 percent for sentences delivered in English [10].

h. Before establishing hierarchical attention, the research advises creating a paragraph-parallel corpus based on the Chinese and English editions of the novels. Our encoder and decoder handle words, phrases, and paragraphs at different level using segmented sentences as input. A two-layer transformer is used to specifically capture the context from the source and destination languages. To condition on its own prior hidden states, the output of the model built on the initial transformer is employed as another level of abstraction [11].

i. With a focus on how adjectives behave in noun phrases, this work provides an analysis of adjectival structures in Head-driven Phrase Structure Grammar (HPSG) (NP). The well-known observation that adjectives without complements usually come before the noun and those with complements or post-modifiers usually come after the noun is developed in this passage [12].

4 Algorithms Used For Machine Translation

In terms of the algorithms utilized, machine translation is also being tried, and as will be mentioned below, there has been a significant amount of success in terms of translation quality.

e.

a. For the error analysis of SaHiT — A Statistical Sanskrit-Hindi Translator, a statistical Sanskrit to Hindi MTS, was proposed for error analysis. The MT Hub platform was used to build the corpus and train it. In the error report produced by the MT Hub System and during the two phases of training for the BLEU, a score of 41% was reported [13].

b. EtranS, a proposed machine translation system for translating English into Sanskrit, aims to enhance translation quality. Additional modules were created, including (a) the Parse Module, where parsing took place following the generation of analysis tokens and analysis of grammar and syntax. (b) The The Generator Module maps semantic data and generates results based on the mapping. The system claimed a 90% accuracy rate for mapping small, large, and extra-large sentences.

c. The use of machine translation to translate English into Sanskrit was suggested. The work was separated into four modules: lexical parser, semantic mapper, translator, and composer. This paper will address the first two modules. A lexical parser was created for POS tag data and its dependencies. This parser produces three different rules: the equality rule, the synonym rule, and the antonym rule.

d. After creating tokens and using dependence to parse them, a tree was constructed and a mapping between English and Sanskrit phrases was finished [14].

Table 2: Summary of various ML Algorithms & Approaches for various Indian Languages with their Success rate %

S.No.	Algorithm/Approaches	Purpose	Year	Success rate in %
1	a) Decision Tree method b) Support Vector Machine	Four methods are applied for the disambiguation in Bengali	2021	DT method-63.84% SVM method-76.9%

	method c) Artificial Neural Network metho d) Naïve Bayes method			ANN method-76.23% NB method-80.23%
2	Naive Bayes Method with syntactic and semantic features	Syntactic as well as semantic features are extracted for the WSD in Assamese language	2019	91.11%
3	a) Naive Bayes Method b) Decision List Classifier c) Maximum Entropy Classifier d) Lazy Boosting	The method is applied on both small and large data sets Apply the algorithm on Spanish Test data Maximum Entropy approach is used for example sentences with rich feature sets Algorithm is applied on the data sets for both the fine grained and coarse grained levels	2019	Large dataset-92.3% Small dataset- 66.4% for verbs and 72.7% for nouns. Simple ambiguities-99% accuracy most difficult ambiguities-90% accuracy Results are more accurate in comparing with the baseline Fine grained- 61.51% accuracy Coarse grained -69% accuracy
4	Naive Bayes Method	Both the models i.e. collocation model and Bag-of-words model is applied for the WSD process for the Punjabi language	2018	81%-89% for both the models
5	Genetic Algorithm	WSD for Gujarati language	2017	Satisfactory
6	Naive Bayes Method	Compare two methods decision tree and NB method and predict the accuracy of NB classification method is better	2016	NB method-62.86% Decision Tree method-45.14%
7	Context Similarity Unsupervised Approach	Apply for the WSD process for the Malayalam language	2016	72%
8	Naive Bayes Method	Used for the Bengali WSD	2015	80-85%
9	Machines of Support Vectors (SVM)	Used for the process of word sense disambiguation of Tamil Language	2014	91.60%
10	Graph-based method of unsupervised learning method	WSD technique used for the disambiguation process of Bengali language	2013	60%
11	Genetic Algorithm	Used this algorithm for the disambiguation of Hindi language	2013	91.60%
12	Modified Lesk's Algorithm	For the disambiguation for the Punjabi language	2011	75%
13	Knowledge Based Approach	Used for the disambiguation process of Malayalam language	2010	81.30%

5 CONCLUSION

Each machine translation mechanism techniques has benefits and drawbacks. The current work is in the direction of building a more accurate algorithm for an English to Sanskrit Machine Translation system. The different machine learning algorithms include various techniques, technologies, and approaches, along with some ambiguity challenges, which have been addressed. Also

find out the accuracy percentage of various algorithms for machine translation.

Declarations

Conflict of interest statement: The authors declare that there is no conflict of interest.

Author's Contribution: All authors equally contributed to the preparation of this manuscript and approved the final.

Ethical Approval and Consent to participate: Not Applicable

Consent for publication: Not applicable

Human and Animal Ethics: Not applicable

Availability of supporting data: This is a review paper. There is no need of dataset.

Funding: Council of Science & Technology sponsored project No CST/D 2475 (Design and Development of Software Tools and Technologies for Translation of Natural Languages). Part of the work presented in the paper was carried out under this research project.

References:

1 Maurya, A. S., Garg, S., & Bahadur, P. (2022). Tools and Techniques for Machine Translation. In Proceedings of Second Doctoral Symposium on Computational Intelligence (pp. 857-867). Springer, Singapore.

2 Promila Bahadur, A.K.Jain and D.S.Chauhan, "EtranS- A Complete Framework for English To Sanskrit Machine Translation" International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Selected Papers from International Conference & Workshop On Emerging Trends In Technology 2012, 2012. http://dx.doi.org/10.14569/SpecialIssue.2012.020107

3 Sreelekha S, Pushpak Bhattacharyya,D. Malathi,Statistical vs. Rule-Based Machine Translation: A Comparative Study on Indian Languages January 2018, International Conference on Intelligent Computing and Applications.

4 J.-X. Cao, D.-G. Huang,W. Wang,S.-J. Wang, January 2014 Dalian Ligong Daxue Xuebao/Journal of Dalian University of Technology 54(1):91-99.

5 Marcio Lima Inácio, Helena Caseli,Word Embeddings at Post-Editing February 2020 Computational Processing of the Portuguese Language, 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings .

6 Elaine Uí Dhonnchadha1 , Josef Van Genabith2, Partial Dependency Parsing for Irish 1Centre for Language and Communication Studies, Trinity College, Dublin 2, Ireland. 2Centre for Next Generation Localisation, Dublin City University, Glasnevin, Dublin

7 Sitender & Seema Bawa, Sanskrit to universal networking language EnConverter system based on deep learning and context-free grammar, Multimedia Systems (2020) Metrics.

8 Ngoc-Bich Le,Xuan-Quy Dao,My-Thanh Nguyen Thi,Design of Text and Voice Machine Translation Tool for Presentations April 2021Conference: 13th Asian Conference on Intelligent Information and Database Systems At: Phuket Thailand .

9 Ariana López-Pereira, Neural Machine Translation and Statistical Machine Translation: Perception and Productivity, December 2019, Revista Tradumàtica.

10 ShreeHarsh Attri T. V. Prasad G. Ramakrishna Computer Science • 21(3) 2020 https://doi.org/10.7494/csci.2020.21.3.3624 HiPHET: Hybrid approach for translating mixed code language (Hinglish) to pure languages (Hindi and English).

11 Yuqi Zhang and Gongshen Liu ,Paragraph-Parallel based Neural

Machine Translation Model with Hierarchical Attention Yuqi Zhang, Gongshen Liu * School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai, Shanghai, 200240, China cici--q@sjtu.edu.cn, lgshen@sjtu.edu.cn, : 2020 J. Phys.: Conf. Ser. 1453 012006.

12 Doug Arnold Louisa Sadler ,Noun-Modifying Adjectives in HPSG ,Department of Language and Linguistics, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK louisa@essex.ac.uk doug@essex.ac.uk

13 Pandey and Jha, Error Analysis of SaHiT - A Statistical Sanskrit-Hindi Translator," 2016.

14 Barkade et al, English to Sanskrit Machine Translation Semantic Mapper.

15 Tapaswi and Jain Morphological and Lexical Analysis of the Sanskrit Sentences, , 2011

16 Tapaswi et al ,Parsing Sanskrit Sentences using Lexical Functional Grammar,2012