# Various Approaches of Machine Translation for Marathi to English Language

*Nilesh* Shirsath[1,3,*], *Aniruddha* Velankar[2,**], and *Ranjeet* Patil[3,***] *Dr.Shilpa* Shinde[3,****]

[1]Department of Computer Engineering, Ramrao Adik Institute of Technology, India

**Abstract.** Machine Translation (MT) is a generic term for computerised systems that generate translations from one natural language to another, with or without human intervention. Text may be used to examine knowledge, and turning that information into pictures helps people to communicate and acquire information.There seems to be a lot of work conducted on translating English to Hindi, Tamil, Bangla and other languages. The important parts of translation are to provide translated sentences with correct words and proper grammar. There has been a comprehensive review of 10 primary publications used in research. Two separate approaches are proposed, one uses rule based approach and other uses neural-machine translation approach to translate basic Marathi phrases to English. While designed primarily for Marathi-English language pairs, the design can be applied to other language pairs with a similar structure.

## 1 Introduction

Machine Translation (MT) is a common name for computerized systems which are responsible for generating, with or without human assistance, translations from one natural language into another.It is part of Natural Language Processing (NLP) where translation from the source language to the target language is conducted, preserving the same meaning of the phrase To help them make text and speech into another language, humans can use Machine Translation Systems. The program can run without any human intervention. The conventional approach is achieved for the MT to translate large quantities of knowledge involving terms that could not be interpreted. The MT performance level can differ considerably , MT programs need ”training to improve the quality of the outcome in the relevant domain and language pair.

As MT can be classified into different categories

- Rule-based Systems: uses a combination of language and grammar rules
- Statistical Systems: learn to translate by analyzing large amounts of data
- Neural Machine Translations (NMT): learn to translate through one large neural network (multiple processing devices modeled on brain)

Different MT providers like Google translator, Yandex Translator etc. provide some translation tools with ethics to customer like:

- Confidentiality – There is no confidentiality in the content translated by online MT platforms such as Google and Microsoft translators. Translated data is stored by the owners of the platform and may later be reused.
- Notifying the Client about MT Use - Whether a translation company should notify customers about the use of MT for their projects is a point of debate in the industry. Many are in pursuit of informing the customer of the use of MT and others may not disclose the use of MT. If you have questions about MT use, be sure to ask your provider.

### 1.1 Objective

The main Objective of the project is to implement a tool which translates Marathi sentences to English without changing the meaning of the sentence using the Rule Based and Neural Based System

### 1.2 Motivation

As the steady progress in the field of technology, the Internet’s growth has also increased at a tremendous rate. With globalization, the official language of the globe has been English. In Marathi literature, there are approximately 71 million Marathi speaking individuals and various works. However, Marathi language is comprehensible for a very small group of people so a system is proposed, which takes the input Marathi sentence and translates it into English which is an understandable language.

## 2 RELATED WORK

A survey of the research done for data summarization and the currently existing systems, give the following results.

A. **Transmuter: An Approach to Rule-based English to Marathi Machine Translation.[1]**

*e-mail: Nileshshirsath2389@gmail.com
**e-mail: aniruddhav25@gmail.com
***e-mail: patilranjeet3699@gmail.com
****e-mail: shilpa.shinde@rait.ac.in

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

The basic method used to implement this framework is Rule Based Machine Translation.The model is built on a generalised approach based on the categories/domains to which a term belongs.The proper spelling of words The target language was created based on the parse tree's basic traversals.The architecture is partly applied in the context of a Machine Translation method.

A generalized approach based on the categories/domains to which a word belongs is the model.In order to achieve better quality translations, the number of rules formed is high for the target language generation.The consistency of the translation of this method depends on The size of the knowledge base of grammar. If the size exceeds this threshold, due to contradictory laws, the consistency can decrease.

B. **Script Translation System For Devanagari To English. [2]**

The proposed scheme will translate more than one Devanagari word to English using the rule-based approach. This is accomplished by understanding the different parts of speech in Marathi phrases. In a bilingual dictionary, tokenizing and describing each word's English meaning and obtaining a clear translation.

The machine translation based on the rule includes generating a number of rules and handling their exceptions as well.Compared to dictionary-based methods that include word-to-word translations, rule-based machine translation provides better translation quality.Given the number of laws to be used in the scheme, flawless translations for each and every sentence will not be done.

C. **Part of Speech Tagger for Marathi Language.[3]**

The rule-based element of the speech tagger, which uses a set of handwritten rules to apply words to all potential tags.The system uses a morphological analyzer to identify the root word and compares it to the corpus to assign appropriate tags.Where an expression has more than one suffix, grammar rules are used to reduce ambiguity.Dictionaries are required in order to assign appropriate tags to each expression.

The basic standard of useful instructions aids in avoiding ambiguity.Because of the lack of corpus for statistical analysis, POS tagging is difficult for the Marathi language. When there is a broad range of meaningful rules to prevent disambiguation, the rules-based POS tagger will achieve greater accuracy.Additional meaningful rules need to be provided to improve the performance of the system

D. **Inflection Rules for English to Marathi Translations[4]**

When it comes to getting the right translation, inflection is crucial.Inflection is the process of changing the meaning of a word by adding the proper suffix to it according to the structure of the sentence.The implementation of the Inflection for English to Marathi Translation is presented in this paper.

The inflection of nouns, pronouns, verbs, and adjectives in a sentence is determined by the other words and their qualities. The rules for inflecting the above Parts-of-Speech are presented in this document.

E. **Marathi to English Neural Machine Translation with Near perfect corpus and transformers [5]**

To translate Marathi sentences into English, the device uses the Neural Machine Translation (NMT) method.It focuses on transformer based architecture.All of the transformers' findings were equivalent to Google Cloud API-V2. In MAE and RMSE, it also outperformed Google API. This study suggests that this is the case.

From the results and examples it is observed that the proposed transformer-based model was able to outperform Google Translation with limited but almost correct parallel corpus.The system produced satisfactory results by making use of word piece tokenizer but in order to improve the performance sentence piece tokenizer must be implemented and the results need to be compared.

F. **Challenges in Rule based machine translation from Marathi to English [6]**

In the field of machine translation, translation divergence is a difficult problem to solve. For proper understanding and identification of divergence issues in machine translation, a thorough investigation is needed. It's a time-consuming process to use Rule Based Machine Translation. The development of a large number of laws necessitates a great deal of human effort.To accomplish these translations, a set of rules must be created.

They explained the different forms of divergence patterns in the Marathi and English language pair in this article. In addition, these divergence trends must be identified and classified.This method's consistency is determined by the scale of the grammatical information base. The scale and depth of the information base grows as more exceptional cases are treated. When the scale of the information base grows, so does the precision, up to a certain point.If the size exceeds a certain threshold, the accuracy can suffer as a result of contradictory laws. As a result, a scheme relying on this method must strike a balance between precision and the number of exceptions it can accommodate.

G. **A survey on LSTM memristive neural network architectures and application [7]**

The first generation of machine translation methods used dictionary-based methods to do word-to-word translations. Its flaws prompted the development of the second generation, which used rule-based and transfer-based techniques.

It has been discovered that rule-based machine translation necessitates the development of a large number of rules as well as the treatment of their exceptions. The system is feasible up to a point, but this approach would have higher translation accuracy.The emphasis of this paper is on Marathi to English translation based on rules.

The Marathi to English translation system will aid in the automation of the process of translating documents and scripts, as well as the reduction of manual translation work.In some sentence translations, there may be some disambiguation. The translation rules, on the other hand, will be framed in such a way that generic sentences or sentences from other domains will be translated.

H. **A Baseline Neural Machine Translation System for Indian Languages [8]**

The research provides a Neural Machine Translation system for Indian languages that is both simple and effective.It establishes a firm baseline for future research by demonstrating the viability of numerous language pairs.

Even if there aren't enough resources, this method, which uses cutting-edge machine learning techniques, produces competitive outcomes for many language pairs.They investigate the multilingual teaching scenario for the Indian language, employing a variety of tried-and-true strategies to get competitive results.These tasks entail calculating embeddings for later classification or sentiment analysis tasks.Indian languages will benefit from the ability to transfer learning from well-tested embeddings, such as English.

I. **On the Properties of Neural Machine Translation: Encoder-Decoder Approaches [9]**

An encoder and a decoder are frequently used in neural machine translation models.From a variable-length input text, the encoder extracts a fixed-length representation, from which the decoder creates an accurate translation.

This paper analyses the properties of neural machine translation using two models: RNN Encoder–Decoder and a newly designed gated recursive convolutional neural network in this research.They show that neural machine translation performs rather well on short phrases with few unknown terms, but that its performance rapidly degrades as the sentence length grows.It also demonstrates that the suggested gated recursive convolutional net-work automatically learns the grammatical structure of a phrase

J. **Neural Machine Translation using Recurrent Neural Network [10]**

The constructed system integrates Recurrent Neural Network models to achieve the maximum accuracies in 10 epochs.It was discovered that using many models at the same time resulted in a better model with improved overall accuracy for the system.

According to the findings of the studies, bidirectional Recurrent Neural Networks with encoding algorithms have a higher accuracy than the other Recurrent Neural Network models. Single Recurrent Neural Network models, such as the basic RNN, only RNN with encoding, only bidirectional RNN, and RNN with encoder and decoder, have lower accuracy than RNN models with encoder and decoder.

K. **BLEU: a Method for Automatic Evaluation of Machine Translation [11]**

According to the report, BLEU will speed up the MT RD cycle by allowing researchers to quickly zero in on useful modelling concepts.A new statistical investigation of BLEU's association with human judgement for translation into English from four distinct languages (Arabic, Chinese, French, and Spanish) representing three different language families supports this viewpoint.

The strength of BLEU is that it has a strong correlation with human assessments since it averages out individual sentence judgement errors over a test corpus rather than attempting to discover the exact human judgement for each sentence: quantity leads to quality.Finally, because MT can be thought of as the generation of natural language from a textual environment, the BLEU might be used to assess MT tasks.

L. **Sequence to Sequence Learning with Neural Networks [12]**

On a large-scale MT challenge, it was demonstrated that a big deep LSTM with a limited vocabulary and essentially no assumptions about problem structure can outperform a traditional SMT-based system with an unlimited vocabulary. Given the success of the simple LSTM technique on MT, it should work well on a variety of other sequence learning problems as long as there is adequate training data.

It was determined that finding a problem encoding with the highest number of short-term dependencies is critical since it makes the learning problem considerably easier.They were unable to train a standard RNN on the non-reversed translation problem using this method, but they believed that when the source sentences were reversed, a standard RNN should be easily trainable.

### 2.1 Limitations of Existing System

- As Marathi language is less researched and progressed on the grammar and translation front the existing systems fail to incorporate all the rules necessary for translation.
- In the Machine Learning approach only a few translation models are implemented which are not providing satisfactory and correct translation in the testing phases due to the lack of the data.

# 3 METHODOLOGY

## 3.1 Proposed Work

### *3.1.1 Rule Based Machine Translation:-*

A system is proposed using rule-based machine translation which follows a sequential procedure of taking input Marathi sentences followed by the machine translation operations such as Tokenization, Tagging etc. Which on the successful implementation tries to predict the most accurate translation of the given sentence.

### *3.1.2 Neural Machine Translation:-*

An Encoder Decoder model will used, which helps to analyzes corpus of data and by understanding pattern between them and for generate the translations.

## 3.2 Techniques

### *3.2.1 Rule Based Machine Translation:-*

Rule-based translation mostly depends on different built-in linguistic rules and millions of bilingual dictionaries for each pair. Translations are done on vast and revolutionary linguistic rules. Automatic translation systems are focused on linguistic knowledge about the source and target languages, essentially derived from dictionaries and grammars (unilingual, bilingualor multilingual) covering the key semantic, morphological and syntactic regularities of each language.

An RBMT method produces output sentences (in some target language) for input sentences(in some source language) based on morphological, syntactic, and semantic study of both the source language and the target involved in a specific task of translation.

### *3.2.2 Neural Machine Translation:-*

In NMT model, a single system can be trained directly on source and target text. Unlike other system NMT works cohesively to maximize its performance and it also used vector representation for words and internal state. The NMT uses a bidirectional recurrent neural network,also called an encoder, to process a source sentence into vectors for a second recurrent neural network, called the decoder, to predict words in the target language. This process, while differing from phrase-based models in method, prove to be comparable in speed and accuracy

## 3.3 Design of the System

### *3.3.1 Rule Based Machine Translation:-*

1. **Source Text**: Takes Marathi sentences as an input text.
2. **Tokenization**: Sentence is broken down into token words.

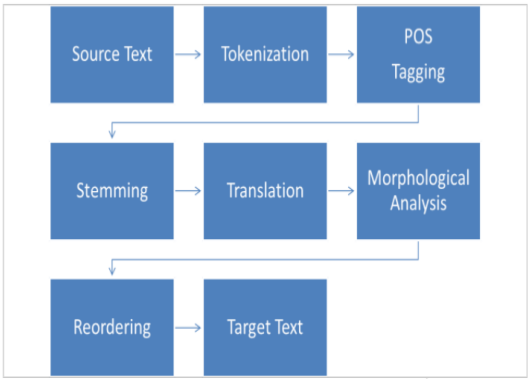


**Figure 1.** System Design of Rule Based Machine Translation

3. **POS tagging:** Part of speech of each token is determined.
4. **Stemming:** From the token root word is stemmed out for translation.
5. **Translation:** Translation of Marathi words is found using a bilingual dictionary
6. **Morphological Generation:** Grammatically correct words are generated according to the suffixes.
7. **Sentence Reordering**: Morphologically Generated words are reordered according to the written grammar rules.
8. **Target Text**: Translated English Sentence is obtained.

### *3.3.2 Neural Machine Translation:-*

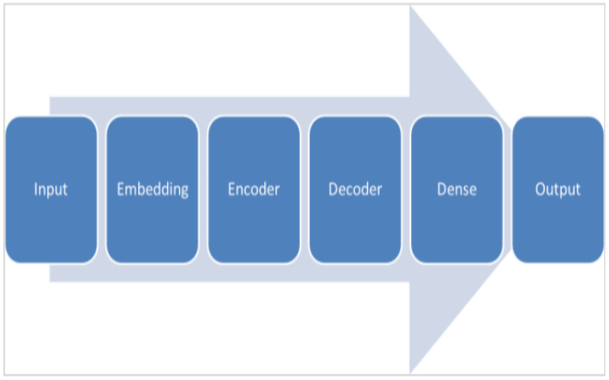


**Figure 2.** System Design of Neural Machine Translation

1. **Input**:Source sentence which is to be translated into an integer encoded form.
2. **Embedding:**In this layer we will map each integer encoded word to an vector which will act as input for next layer.There are various word embedding techniques which map(embed) a word into a fixed length vector.

3. **Encoder:**The encoder reads the input sequence and summarises the data in what are known as internal state vectors (in case of LSTM these are called as the hidden state and cell state vectors). The encoder outputs are discarded, leaving only the internal states.

4. **Decoder:**The initial states of the Decoder LSTM are set to the final states of the Encoder LSTM. The decoder begins generating the output sequence using these initial states.

5. **Dense:**It's used in this network explicitly to adjust the LSTM layer's dimensions to the one we like. In this case, the network's output is expected to be a probability distribution over all of the vocabulary terms.

6. **Output:**Predicted translation of input sentence into one hot encoded form.

## 4 RESULT AND ANALYSIS

After analyzing the papers and implementing RBMT approach we were able to extract a few linguistic rules related to Marathi language which are helpful for translation of sentences.After studying many Marathi sentences we realized the derived words are formed by adding suffixes in most of the cases.We collected some of the common suffixes and separated on the basis of POS tags for the simplification purpose.System is able to translate the sentences having the structures and rules which were analyzed but the system struggles to translate the sentences which occur having the different and complex structure.

On the other hand NMT achieves decent translation whose accuracy can be judged by BLEU score.It does not requires dedicated POS tagging or stemming since the model learns it while training.To achieve great quality of translation using NMT large corpus of cleaned data is required and also the large computational power is necessary to train the model.NMT is able to use algorithms to learn linguistic rules on its own from statistical models.The biggest benefit to NMT is its speed and quality.NMT makes faster translations than the Rule Based method and has the ability to create higher quality output.

| | |
|---|---|
| BLEU-1 | 0.865143 |
| BLEU-2 | 0.800385 |
| BLEU-3 | 0.758851 |
| BLEU-4 | 0.660873 |

Accuracy of model on training data:

| | |
|---|---|
| BLEU-1 | 0.623246 |
| BLEU-2 | 0.507332 |
| BLEU-3 | 0.455952 |
| BLEU-4 | 0.341381 |

Accuracy of model on testing data:

## 5 CONCLUSION AND FUTURE WORKS

The RBMT approach shows how exactly translation works and give almost perfect translation of given sentence if rule for it is written correctly and development of a such large number of laws required great human effort.As the rule increases the consistency of result will decrease because of treatment of some exceptions but it produces the result correctly and it's also depends on scale of grammatical information base.The system is feasible up to a point, but this approach would have higher translation accuracy.

About NMT approach its not necessary to have deep linguistic knowledge, but we required good amount of cleaned data so the model can train properly and develop its own rule in order to give proper translation.The BLEU-4 score of about 0.7286, provides an upper bound on what we might expect the efficiency from this model. A BLEU-4 score of about 0.3510 was achieved, providing a baseline skill to improve upon with further changes and improvements to the model.

The accuracy of Rule Based Translation may be improved by developing better rules, which may be achieved through more language research and assistance from linguists or language experts. A huge corpus of high-quality labelled data for the Marathi language can help enhance machine translation utilizing a neural-based methodology.

## 6 ACKNOWLEDGMENT

The authors thank the many people who have done lots of help for us and provided useful guidance for our paper.

## References

[1] G V Garje,G K Kharate,Harshad Kulkarni,"Transmuter: An Approach to Rule-based English to Marathi Machine Translation",(2014).

[2] Jayshri A.Todase and Sushama Shelke ,"Script Translation System For Devnagari To English",(2018).

[3] Sharvari Govilkar, Bakal J W and Shubhangi Rathod,"Part of Speech Tagger for Marathi Language",(2015)

[4] Inflection Rules for English to Marathi Translation , IJCSMC, Vol. 2, Issue. 4, April 2013, pg.7 – 18

[5] Swapnil Ashok Jadhav,"Marathi To English Neural Machine Translation With Near Perfect Corpus And Transformers",(2020).

[6] Namrata G Kharate and Dr. Varsha H. Patil,"Challenges in rule based machine translation from marathi to english",(2019).

[7] G V Garje,Adesh Gupta,Aishwarya Desai,Nikhil Mehta,Apurva Ravetkar,"Marathi to English Machine Translation for Simple Sentences",(2014).

[8] A Baseline Neural Machine Translation System for Indian Languages,Jerin Philip, Vinay P. Namboodiri, C.V. Jawahar",(2019).

[9] Cho, Kyunghyun van Merriënboer, Bart Bahdanau, Dzmitry Bengio, Y.. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. 10.3115/v1/W14-4012

[10] Neural Machine Translation using Recurrent Neural Network,International Journal of Engineer-

ing and Advanced Technology (IJEAT)ISSN: 2249 –8958,Volume-9 Issue-4, April 2020

[11] BLEU: a Method for Automatic Evaluation of Machine Translation , Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,(2002).

[12] Sequence to Sequence Learning with Neural Networks,Ilya Sutskever, Oriol Vinyals, Quoc V. Le",(2014).