Neural Machine Translation from English to Marathi Using Various Techniques

Yash Jhaveri, Akshath Mahajan, Aditya Thaker, Tanay Gandhi, and Chetashri Bhadane

Abstract Machine translation (MT) is a term used to describe computerized systems that generate translations from one linguistic communication to another, either with or without the need of humans. Text can be used to evaluate knowledge and converting that information to visuals can help in communication and information acquisition. There have been limited attempts to analyze the performance of state-of-the-art NMT algorithms on Indian languages, with a significant number of attempts in translating English to Hindi, Tamil and Bangla. The paper explores alternative strategies for dealing with low-resource hassle in neural machine translation (NMT), with a particular focus on the English-Marathi NMT pair. To provide high-quality translations, NMT algorithms involve a large number of parallel corpora. In order to tackle the low-resource dilemma, NMT models have been trained, along with transformers and attention models, as well as try hands-on sequence-to-sequence models. The data has been trained for sentence limit of 50 words and then fine-tune the default parameters of these NMT models to obtain the most optimum results and translations.

Keywords English · Marathi · Natural language processing · Neural machine translation · Rule-based machine translation · Transformers · Attention models · Seq2Seq models

1 Introduction

Communication has been an integral part of human life ever since the beginning of time. As per Census of India, 2011, Marathi is the third most frequently spoken language in India and ranks 15th in the world in terms of combined primary

Yash Jhaveri, Akshath Mahajan, Aditya Thaker, Tanay Gandhi have contributed equally.

Y. Jhaveri (✉) · A. Mahajan · A. Thaker · T. Gandhi · C. Bhadane
DJ Sanghvi College of Engineering, Mumbai, India
e-mail: yashj65@gmail.com

Y. Singh et al. (eds.), *Proceedings of International Conference on Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 1011,
https://doi.org/10.1007/978-981-99-0601-7_3

and secondary speakers [1]. It is spoken by roughly 83 million of the world's 7 billion people [2]. In today's modern world, majority research and other materials are written in English, which is ubiquitously recognized and valued. Existing Marathi documents must be translated into English for them to be universally used. However, manual translation is time consuming and expensive, necessitating the development of an automated translation system capable of performing the task efficiently. Furthermore, there hasn’t been much advancement in translating Indian languages. English is a Subject-Verb-Object language, whereas Marathi is Subject-Object-Verb with relatively free word order. Consequently, translating it is a difficult task.

MT refers to computerized systems that generate translations from one linguistic communication to another, either with or while not human involvement. It is a subset of natural language processing (NLP) wherein translation from the source language to the target language is undertaken while conserving the same meaning of the phrase. Furthermore, neural machine translation (NMT) has achieved tremendous progress in recent years in terms of enhancing machine translation quality (Cheng et al. [3]; Hieber et al. [4]). The encoder and decoder, which are commonly based on comparable neural networks of different sorts, such as recurrent neural networks (Sutskever et al. [5]; Bahdanau et al. [6]; Chen et al. [7]), and more recently on transformer networks, make up NMT as an end-to-end sequence learning framework (Vaswani et al. [8]).

The proposed work seeks to improve the English-to-Marathi translations and vice-versa and try to mitigate the low-resource problem. The paper proposes a method to enact translations using the paradigmatic NMT models accompanied by the state-of-the-art models like Sequence2Sequence models, attention and transformer models taking into consideration models like SMT along with rule-based learning as the baselines.

2 Literature Survey

In recent years, NMT has made significant progress in improving machine translation quality. Google Translate [9], Bing Translator [10] and Yandex Translator [11] are some of the most popular free online translators, with Google Translator [9] being one of the most popular locations for machine translation.

The state-of-the-art approaches for machine translation, which includes rule-based machine translation and NMT, have been widely used [12–16]. Rule-based MT primarily connects the structure of given input sentences to the structure of desired output sentences, ensuring that their distinctive meaning is preserved. Shirsath et al. [12] offer a system to translate simple Marathi phrases to English utilizing a rule-based method and a NMT approach, with a maximum BLEU score of roughly 62.3 in the testing set. Garje et al. [13] use a rule-based approach to develop a system for translating simple assertive and interrogative Marathi utterances into matching English sentences. Due to the lack of a large corpus for translation, Govilkar et al. [14] used rule-based techniques to translate only the components of speech for the sentence. The proposed system uses a morphological analyzer to locate root words and then compare the root word to the corpus to assign an appropriate tag. If a word contains more than one tag, ambiguity can be eliminated using grammatical rules. Garje et al. [15] present an online parts of speech (POS) tagger and a rule-based system for translating short Marathi utterances to English sentences. Garje et al. [16] primarily focus on the grammar structure of the target language in order to produce better and smoother translations and employ a rule-based approach to translate sentences, primarily for the English–Marathi pair, with a maximum BLEU score of 44.29. Banerjee et al. [17] specifically focus on the case of English–Marathi NMT and enhance parallel corpora with the help of transfer learning to ameliorate the low-resource challenge. Techniques such as phrase table injection (PTI) have been employed and for augmenting parallel data, pivoting and multilingual embeddings to leverage transfer learning, back-translation and mixing of language corpora are used.

Jadhav [18] has proposed a system where a range of neural machine Marathi translators were trained and compared to BERT-tokenizer-trained English translators. The sequence-to-sequence library Fairseq created by Facebook [19] has been used to train and deduce with the translation model.

In contrast with the NMT model, there has been a quite significant upscale in other models that can be used along with the state-of-the-art NMT models for MT. Vaswani et al. [8] have deduced that when compared to conventional recurrent neural network (RNN)-based techniques, the transformer model provides substantial enhancements in translation quality which was proposed by Bahdanau et al. [6], Cho et al. [20] and Sutskever et al. [5]. Self-attention and absence of recurrent layers can be used alongside state-of-the-art NMT models that enable training quicker and a better performance in the case of absence of a huge corpus for translation.

3 Research Gap

Google Translate [9] mainly uses statistical MT models, parameters of which are obtained through analysis of bilingual text corpora, i.e., sentences that have poor quality text translations. Furthermore, BLEU score of the translation received for sentences less than 15 words is 55.1, and above 15 words is 28.6.

The rule-based technique employed by [12–16] is now obsolete and is being replaced by transformers, deep learning models that employ the mechanism of self-attention. Furthermore, Shirsath et al. [12] have provided a maximum BLEU score of about 62.3 in the testing set using rule-based techniques, whereas the paper has achieved a maximum BLEU score of about 65.29 using the proposed methodology. Govilkar et al. [14] translated only the parts of speech for the sentence using rule based techniques. In order to increase the system's performance, extra meaningful rules must be added. Garje et al. [16] have also used rule-based techniques for translation but have provided a maximum BLEU score of around 49, whereas the paper has achieved a maximum BLEU score of about 65.29 using the proposed methodology. Moreover, the problem with rule-based learning lies with exploring with the incomprehensible grammar, which is on the other hand eliminated by the approach presented by the paper. Newer techniques such as phrase table injection (PTI), back-translation and mixing of language corpora have been applied by Banerjee et al. [17], yet have failed to achieve an adequate BLEU score having used a huge corpus of around 2.5 lakh sentences. From the results from the proposed system of Jadhav [18], it can be observed that the proposed transformer-based model can outperform Google Translation for sentence length up to 15 words but not more than 15 words. This paper, on the other hand, focuses on sentences more than 15 words length and tries to model accurate predictions.

4 Methodology

4.1 Data Used and Data Preprocessing

The dataset used is the parallel corpus data from "https://www.manythings.org/anki/". Processing of around 44486 samples from the dataset has been carried out. The sentences were almost clean, but some preprocessing was required. The special characters, extra spaces, quotation marks and digits in the sentences were removed, and the sentences were lowercase. The paper compares the performance of language translation by restricting the length of the sentences to 15 and 50. The target sentences were prefixed and suffixed by the START and END keywords. The authors padded the shorter sentences after the sentence using the Keras pad_sequences method. The dataset was tokenized using the TensorFlow dataset's SubWordTextEncoder (Table 1).

4.2 Model Architecture

Statistical MT [21] is one of the most widely used techniques in which conditional probabilities are calculated using a bilingual corpus, which is used to reach the

Table 1 Dataset examples

English	Marathi
Could you get me some tea?	मला थोडा चहा आणून देशील का?
I'm doing what I can	मी जे करू शकतो ते मी करतोय⌧
Do you really live alone?	तुम्ही खरच एकटे राहता का?
Tom was also shot in the leg	टॉमला पायातसुद्धा गोळी लागली⌧
I also like cakes	मला केकसुद्धा आवडतात⌧

most likely translation. As a baseline model, SMT model has been employed to convert English sentences to Marathi. This was achieved through a word-based SMT model, trained by calculating the conditional probabilities of Marathi words given an English word, and using it to translate input sequences token by token. Most translation systems are based on this technique but do not achieve precise translations.

In order to tackle this, newer methods like rule-based MT and NMT had been introduced with the most accurate method being NMT. This method employees NLP concepts and includes models like sequence-to-sequence, attention and transformers.

Sequence-to-sequence. RNNs [22] are a type of artificial neural networks, which were one of the first to be used to work with sequential or time series data. RNNs require that each timestep be provided with the current input as well as the output of the previous timestep. Although it stores context from past data in the sequence, it is also prone to vanishing and exploding gradient problems. LSTMs were introduced to overcome this problem, by maintaining forget, input and output gates within each cell, that controls the amount of data which is stored and propagated through the cell.

Sequence-to-sequence (seq2seq) models [23] are a class of encoder–decoder models that are used to convert sentences in one domain to sentences in another domain. This encoder–decoder architecture comprises the encoder block, the decoder block and context vector.

1. Encoder block: This block consists of a stack RNN layer, preferably with LSTMs cells. The outputs of the encoder block are discarded, as the hidden states of the last LSTM cell are used as a context vector and sent to the decoder block.
2. Decoder block: This block consists of the same architecture as that of the encoder block. It is trained for a language modeling task, in the target language taking only the states of the encoder block as input (Fig. 1).

The image above describes the architecture of the encoder–decoder model. During the training phase of the decoder, teach forcing is used, which feeds the model

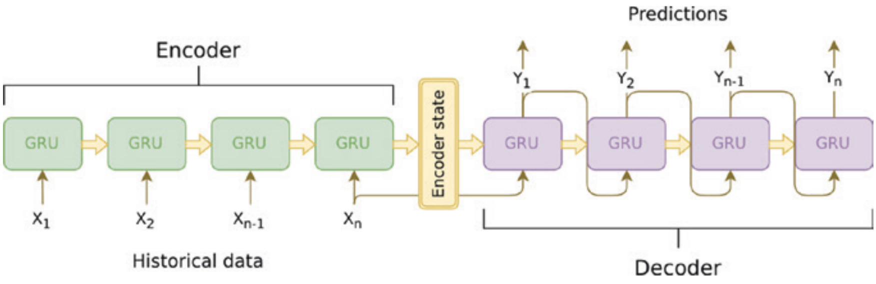

Fig. 1 Seq2Seq [24]

ground truth instead of the output of the previous states. In the testing phase, a <START> token is provided as input to the first cell of the decoder block that marks the start of a sequence, along with the hidden states of the encoder block. The outputs of this cell are used as input to the next cell to make a prediction for the next word. This procedure continues, until the <END> token is generated which marks the end of the sequence. This token is used so that the model can be assured that the sentence translation procedure has finished.

A single RNN layer has been used consisting of LSTM cells for the encoder block and a similar architecture for the decoder block. Embedding layers are used to translate the sentences from words to word vectors before it can be used by the encoder. Another embedding layer is used to convert the outputs of the decoder block into words in target language, after which a softmax function gives a probability distribution over the vocabulary.

Attention. In recent years, NMT problems have found major success using the encoder–decoder framework, which first encodes the source sentence, that is used to generate the translation by selecting tokens from the target vocabulary one at a time. [22, 23]

This paradigm, however, fails on long sentences where the context required to correctly predict the next word might be present at a different position in the sentence which might be forgotten. An attention mechanism is used to refine translation results by focusing on important parts of the source sentences [25] (Fig. 2).

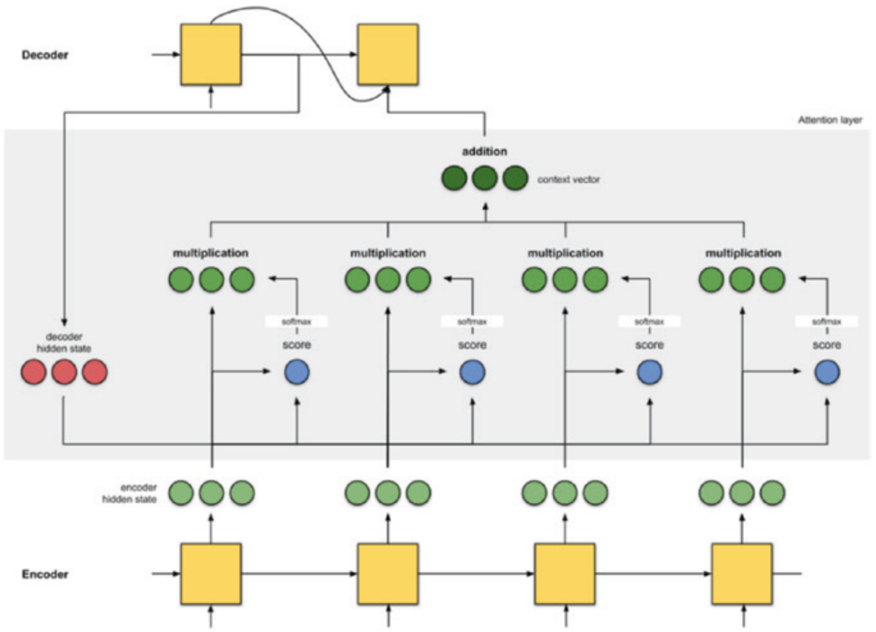

Fig. 2 Illustrated attention [26]

The proposed encoder network consists of three LSTM layers having 500 latent dimensions. On the other hand, the decoder network first has an LSTM that has its initial state set to the encoder state. The attention layer is then introduced that takes the encoder outputs and the outputs from the decoder LSTM. Finally, the outputs from the decoder LSTM and the attention layer are combined and passed through a time-distributed dense layer.

The authors have used the "Teacher Forcing" method to train the network faster. The model was set to train for 40 epochs using the RMSProp optimizer along with sparse categorical cross-entropy loss but observed early stopping after just 22 epochs.

The trained weights are then saved, and an inference model is generated using the encoder and decoder weights to predict and evaluate the translation results. This is done by adding a fully connected softmax layer after the decoder in order to generate a probability distribution over the target vocabulary.

Transformers. The work by Ashish Vaswani et al. [8] proposes a novel method for avoiding recurrence and depending solely on the self-attention mechanism. This new architecture is more precise, parallelizable and faster to train (Fig. 3).

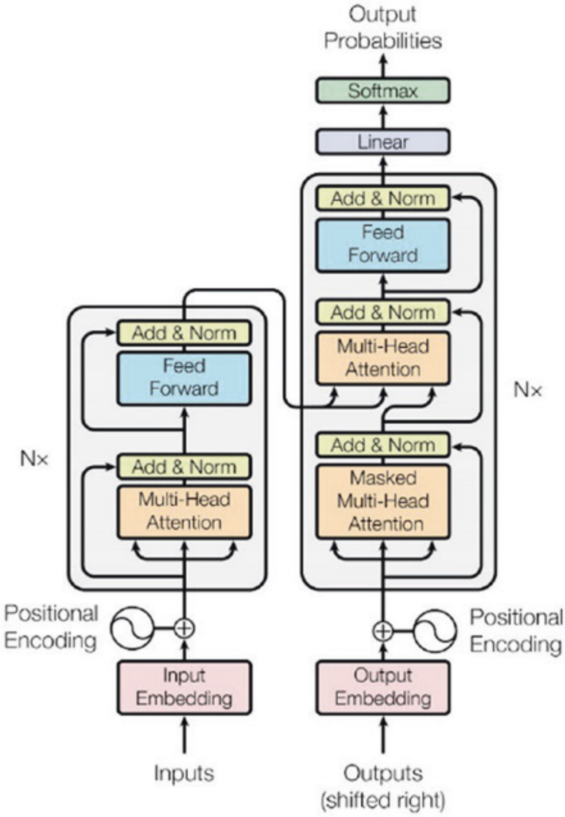

Fig. 3 Transformer architecture [8]

In the transformer model, a stack of six encoders and six decoders is used. The input data is first embedded before it is passed to the encoder or decoder stacks. Because the model lacks recurrence and convolution, the authors injected some information about the relative or absolute positions of the tokens in the sequence to allow the model to use the sequence's order. Positional encoding was added to the input embeddings to achieve this. The positional encodings and embeddings have the same dimension, therefore can be added together.

There are two levels to each encoder. The multi-head attention layer is the initial encoder in the stack through which the embeddings with their positional encoding are passed and subsequently supplied to the feed-forward neural network. The self-attention mechanism uses each input vector in three different ways: the query, the key, and the value. These are transmitted through the self-attention layer, which calculates the self-attention score by taking the dot product of the query and key vectors. To have more stable gradients, this is divided by the square root of the dimensions of key vectors and then supplied to the softmax algorithm to normalize these scores. This softmax score is multiplied by the value vectors, and then the sum of all weighted value vectors is computed. These scores indicate how much attention should be paid to other parts of the input sequence of words in relation to a certain word. Because the self-attention layer is a multi-headed attention layer, the word vectors are broken into a predefined number of chunks and transmitted through various self-attention heads to pay attention to distinct parts of the words. To generate the final matrix, the output of each of these pieces is concatenated and multiplied by the specified weight matrix. This is the final output of the self-attention layer, which is normalized and added to the embedding before being sent to the feed-forward neural network.

5 Results

After experimenting with the number of layers in the model and fine-tuning the hyperparameters of the models used, the paper compares the results of the translations produced using the BLEU score and WER score.

The sacreBLEU score is a metric for assessing the quality of machine translations from one language to another. The link between a machine's output and that of a human is characterized as quality. It was created to evaluate text generated for translation, but it can also be used to evaluate text generated for other natural language processing applications. Its output is often a score between 0 and 100, indicating how close the reference and hypothesis texts are. The higher the value, the better the translations.

Word error rate (WER) computes the minimum edit distance between the human-generated sentence and the machine-predicted sentence. It calculates the number of discrepancies between the projected output and the target transcript by comparing them word by word. The smaller the value, the better the translations.

From the Tables 2, 3, 4, 5, 6, 7 and Fig. 4, it can be observed that the best performing model with respect to SacreBLEU score and WER score metrics is the transformer model, while the worst performing model is SMT. This is so because the transformer model keeps track of the various word positions in the sentences and uses the attention mechanism while the SMT depends upon the probability of the next word which makes it less accurate and reliable.

Table 2 Comparison of various metrics for various models

Metrics	Sequence-to-sequence	Attention	Transformers	SMT
SacreBLEU score	64.49	61.8	65.29	48.22
WER	1.87	4.0	1.55	3.4

Table 3 Translation result

Input	I really didn’t have time
Required	माझ्याकडे खरच वेळ नव्हता⊠
SMT	वेळ खरच तुमच्याकडे मी⊠
Seq2Seq	माझ्याकडे खरच वेळ नव⊠
Attention	माझ्याकडे खरच वेळ नव्हता⊠
Transformer	माझ्याकडे खरच वेळ नव्हता⊠

Table 4 Translation result

Input	Ive already finished reading this book
Required	हे पुस्तक माझं आधीच वाचून झालं आहे⊠
SMT	वाचत आधीच पुस्तक या हे⊠
Seq2Seq	पुस्तक माझं आधीच वाचून झालं⊠
Attention	मी आधीच हे पुस्तक वाचून काढलं आहे⊠
Transformer	माझं आधीच हे पाच वाजलं आहे⊠

Table 5 Translation result

Input	I don’t understand your answer
Required	मला तुमचं उत्तर समजलं नाही⊠
SMT	समजत उत्तर तुमचं मी⊠
Seq2Seq	मचं उत्तर समजलं⊠
Attention	मला तुझं उत्तर समजत नाही⊠
Transformer	मला तुझं उत्तर समजलं नाही⊠

Table 6 Translation result

Input	Did you drink coffee yesterday
Required	काल तू कॉफी प्यायलास का⊠
SMT	के⊠लं तुम्ही पीत कॉफी काल⊠
Seq2Seq	कॉफी प्यायला⊠
Attention	काल तू कॉफी काल के⊠ास का⊠
Transformer	काल तू कॉफी प्यायलीस का

Table 7 Translation result

Input	Whose umbrella is this
Required	ही छत्री कोणाची आहे⊠
SMT	आहे कोणाची या ही⊠
Seq2Seq	छत्री कोणाची⊠
Attention	ही छत्री कोणाची आहे⊠
Transformer	ही छत्री कोणाची आहे⊠

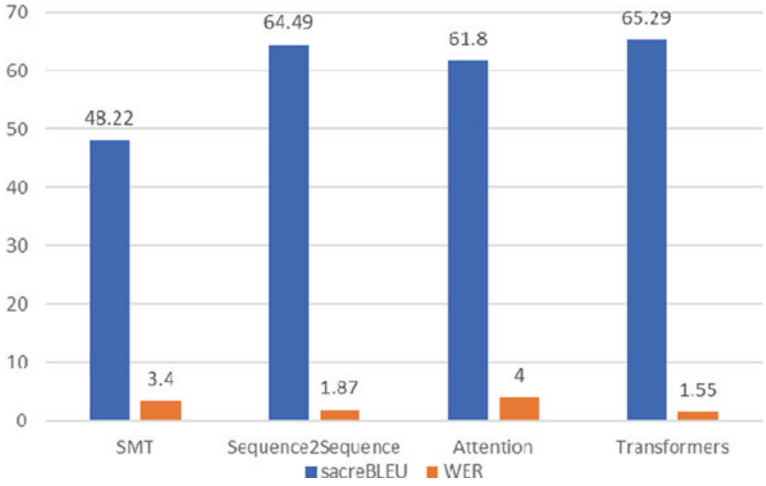

Fig. 4 SMT versus Sequence2Sequence versus attention versus transformer

6 Conclusion

After scrutinizing and implementing different models like Sequence2Sequence, attention models, transformers and SMT, the authors have arrived at the conclusion that after training all mentioned models over a low corpus, the leading fidelity has been obtained by the transformers model. The BLEU Score of about 65.29 and The WER Score of 1.55 state an upper bound on the efficiency of this model. To conclude, the authors did not only mitigate the low-resource problem but also discerned how exactly the translation works and moreover provides almost the exact translations of the given sentence.

References

1. https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers. Last accessed 03 Sept 2022
2. https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf. Last accessed 03 Sept 2022

3. Cheng Y, Xu W, He Z, He W, Wu H, Sun M, Liu Y (2016) Semisupervised learning for neural machine translation. Assoc Comput Linguistics, 1(Long Papers):1965–1974
4. Hieber F, Domhan T, Denkowski M, Vilar D, Sokolov A, Clifton A, Post M (2017) Sockeye: A toolkit for neural machine translation, arXiv preprint, arXiv:1712.05690
5. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. Int Conf Neural Inf Process Syst 2:3104–3112
6. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. Int Conf Learn Represent
7. Chen MX, Firat O, Bapna A, Johnson M, Macherey W, Foster G, Jones L, Parmar N, Schuster M, Chen Z (2018) The best of both worlds: combining recent advances in neural machine translation. Assoc Comput Linguistics 76–86
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 5998–6008
9. https://translate.google.com/. Last accessed 03 Sept 2022
10. https://www.bing.com/translator. Last accessed 03 Sept 2022
11. https://translate.yandex.com/. Last accessed 03 Sept 2022
12. Shirsath N, Velankar A, Patil R, Dr. Shinde S Various approaches of machine translation for Marathi to English language. ICACC-2021 (vol 40)
13. Garje GV, Bansode A, Gandhi S, Kulkarni A (2016) Marathi to English sentence translator for simple assertive and interrogative sentences. Int J Comput Appl 138(5):0975–8887
14. Govilkar S, Bakal JW, Rathod S (2015) Part of speech tagger for Marathi language. Int J Comput Appl 119(18):0975–8887
15. Garje GV, Gupta A, Desai A, Mehta N, Ravetkar A (2014) Marathi to English machine translation for simple sentence. Int J Sci Res (IJSR) 3(11):3166–3168
16. Garje GV, Kharate GK, Eklahare, Kulkarni NH (2014) Transmuter: an approach to rule-based English to Marathi machine translation. Int J Comput Appl 98(21):0975–8887
17. Banerjee A, Jain A, Mhaskar S, Deoghare S, Sehgal A, Bhattacharyya P (2021) Neural machine translation in low-resource setting: a case study in English-Marathi pair. In: Proceedings of the 18th biennial machine translation Summit Virtual USA, August 16–20, 2021, vol 1. MT Research Track
18. Jadhav SA Marathi To English neural machine translation with near perfect corpus and transformers. Available [Online]: https://www.researchgate.net/publication/339526359_Marathi_To_English_Neural_Machine_Translation_With_Near_Perfect_Corpus_And_Transformers
19. https://fairseq.readthedocs.io/en/latest/. Last accessed 03 Sept 2022
20. Cho K, van Merrienboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder-decoder approaches. CoRR, arXiv preprint arXiv:1409.1259
21. Brown PF, Cocke J, Pietra Della SA, Pietra Della VJ, Jelinek F, Lafferty JD, Mercer RL, Roossin PS (1990) A statistical approach to machine translation. Comput Linguistics 16(2):79–85
22. Sherstinsky Alex (2020) Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena 404:132306
23. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst 27
24. https://github.com/sooftware/seq2seq. Last accessed 03 Sept 2022
25. Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025
26. Karim R (2019). Attn: illustrated attention. Towards data science, on medium, January 20. Accessed 15 May 2022