ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 9 (2022) No.3, pp.271-283
DOI:10.33168/LISS.2022.0319

A Study on Approaches to Neural Machine Translation

Basab Nath[1+], Chandrashekhar Kumbhar[1], Bui Thanh Khoa[2]

[1] School of Engineering, ADYPU, India
[2] Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

[+] *basabnath@gmail.com (Corresponding author), mr.kshekhar@gmail.com, buithanhkhoa@iuh.edu.vn*

Abstract. Machine translation is a computer-oriented translation procedure that takes a set of words belonging to a specific human readable language as input and returns another set of words in the desired human readable language as output. The advent of machine translation has benefited sociologists and linguists across the globe. A machine translation model can be, Rule-based, Statistical based, or Neural Network based. Due to the multiple limitations associated with rule-based and statistical-based machine translation models, Neural Machine Translation (NMT) model came into the picture. Machine translation models based on neural networks showed promising results for different human-readable languages, with a rich vocabulary trained with a large corpus. This paper aims to present a literature study on different approaches for different operations in Neural Machine Translation, such as word embedding, attention mechanisms, etc. Also, we emphasized our survey on Neural Machine Translation systems dedicated to Indian languages.

Keywords: NMT, encoder-decoder, RNN.

1. Introduction

Humans communicate among themselves with the help of a mutually understandable language. Communication is essential for sharing knowledge and socializing. Nevertheless, as there are roughly more than 6500 different languages across the globe, language can be a barrier to communication. This is where machine translation comes in. With the help of rich data sets accompanied by appropriate methodologies, machine translation will be able to generate quality translations, which have the potential to surpass human translators. However, in order to attain accuracy to such an extent, multiple factors are to be considered, namely, an abundance of the vocabulary of both source and target languages, comprehension of semantic properties of both source and target languages, and many more. The viable models for machine translation are rule-based, statistical-based, and neural networks-based (Khoa *et al.*, 2021; Kim & Lim, 2022; Koehn & Knowles, 2017). The rule-based methods are based on maintaining huge dictionaries containing predefined semantic and syntactical rules of the concerned languages. The statistical-based models use probabilistic functions for statistical computing tables, which comprise the semantic rules learned from the bilingual corpus. The statistical tables may be syntax trees or phrase translation models. The advent of neural machine translation models eliminates the need for feature engineering tasks such as building parse trees, phrase translation models, rule-based dictionaries, etc. In the case of Indian languages, the linguistic resources are low, especially for the languages spoken in the North-East region of India. Our survey focuses on exploring the techniques in Neural Machine Translation, which can be used to deal with issues found in low-resource languages, such as Out Of Vocabulary (OOV) words, rare words, and many more (Sennrich *et al.*, 2015).

Nonetheless, morphological and structural diversity complicates translating English into Indian languages. When English is translated into Indian languages, a unique set of challenges arises in the form of parallel corpora and linguistic variations. The use of the Assamese language on the internet has become significantly more common during the past few years. On the other hand, just as with other Indian languages, there is very few translations system that has been built that would provide users with a speedy and accurate translation of the text. Therefore, a study of low-resource natural machine translation is useful for young academics starting out in this subject and industry practitioners. There are currently surveys on many aspects of natural language processing (such as multilingual translation and domain adaptation). Still, there is not yet a comprehensive survey for natural language processing using low resources.

As a consequence of this, we have constructed the paper in two sections. First, the paper we are presenting here involves conducting a survey on low-resource NMT that is both comprehensive and well-structured in order to fill in this gap. Because much work has not been done in this area, we concluded that the best approach would be to create and test three different neural machine translation models using this language combination.

2. Preliminaries

2.1. Neural machine translation (NMT)

Neural Machine Translation (NMT) is a translation paradigm that use artificial neural networks to translate an input text from one language to another. The NMT model is a many-to-many model that is also known as a sequence-to-sequence model. The input and output sequences of a machine translation system might be of any length, which is why a many-to-many model is utilized. The NMT model is built as an Encoder-Decoder architecture, similar to a sequence-to-sequence model, and is trained on a massive multilingual corpus. This Encoder-Decoder architecture includes an encoder that accepts a sequence from the source language and a decoder that creates the corresponding translated sequence in the destination language. The Encoder accepts a series of word vectors as input and encodes the whole sequence into a vector known as the context vector. The context vector's goal is to succinctly convey a full source phrase. The Decoder takes the context vector as input and creates the proper word vector sequence. The vector representations of the corpus vocabulary are known as word vectors. Encoders and decoders are examples of artificial neural networks (Sennrich *et al.,* 2015).

Recurrent Neural Networks (RNNs): are neural network topologies that function well with time series and sequential data. RNNs, unlike feed-forward neural networks, calculate their outputs based not only on the current input but also on prior input sequences. These neural networks have memory states which are responsible for storing previous dependencies required for predicting sequences (Christalin *et al.*, 2022; Shah & Bakarola, 2019).

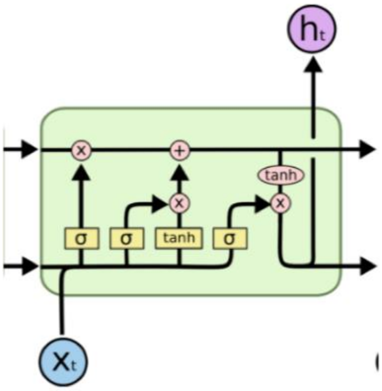
Long Short-Term Memory (LSTMs): Standard RNNs have an issue with Vanishing and Exploding gradients, which prevents them from learning from longer sequences. So, there came along multiple variants of RNNs which elevate these issues to an extent. LSTM is one of them. LSTMs tend to perform better than the standard RNNs when it comes to retaining longer sequences. They use three different gates called the input, output and forget gates respectively. LSTMs maintain two states such as the cell state and the hidden state (Christalin *et al.*, 2022).

2.2. Gated recurrent units (GRUs)

GRU is a simplification to the LSTM architecture. LSTMs as we have seen, are able to perform very well with longer sequences as compared to standard RNNs. But it comes with a cost which is, LSTMs take longer time to train and occupy more computational resources. As a result, in GRU, the cell state and hidden state in LSTM are integrated into a single hidden state. The number of gates is also decreased to two, which are the reset gate and the update gate.

Reset Gate. This gate determines how much of previous hidden state information is used to compute the current hidden state. This gate decides what memory to remember and what memory to forget.

Update Gate. This gate decides whether to calculate the current hidden state using prior hidden state information or the current input.



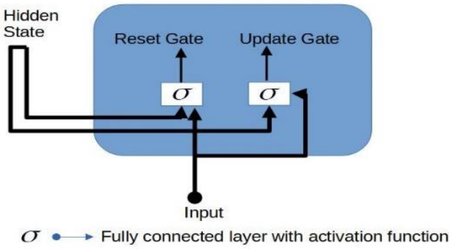


Fig1(a). Pictorial Representation of a LSTM Fig1(b). Pictorial Representation of a GRU

2.3. Encoder-decoder architecture with attention mechanism

The typical Encoder-Decoder design has a bottleneck in that the full input sequence information is compressed into a fixed length vector called the context vector. The decoder uses this context vector to anticipate the target sequence. This limited length context vector cannot maintain longer sequences, resulting in poor translation. As a result, an extra layer in the Encoder-Decoder architecture known as the attention layer enters the picture. The context vector will have access to the whole state history of the encoder for each decoding time step thanks to this attention layer. This provides the decoder with a comprehensive representation of the input sequence. The attention layer enables the decoder to pay varying levels of attention to various input words at different stages of decoding.

Let x = [x1,x2,x3,....,xn] be the input sequence and y = [y1,y2,y3,.....,ym] be the output sequence and $h_t = f(x_t, h_{t-1})$ denote the hidden state of the encoder, for t = 1,2,...,n.

Let the decoder hidden state, $d_t = f(d_{t-1}, y_{t-1}, c_t)$for the output word at position t = 1,2,....,m. Here, f can be any instance of a recurrent neural network.

The context vector c_t , is the weighted sum of the encoder hidden states. The context vector c_t is given as follows: $c_t = \sum_{i=1}^{n} \alpha_{t,i}\, h_i$, for output y_t

In the attention model proposed by Bahdanau et al. **[**Cho et al, 2014**],** the variable sized alignment vector, $\alpha_{t,i}$ given as,

$\alpha_{t,i} = \frac{exp(e_{ti})}{\sum_{k=1}^{n} exp(e_{tk})}$, where, $e_{ti} = a(d_{t-1}, h_i)$ is the alignment model.

The alignment model provides a scoreα_(t,i), to the pair of input at position i and output at location t depending on how well they match. α_(t,i)is the vector containing weights which determine how much of each encoder hidden states should be considered for each output. This alignment model 'a' is trained alongside all of the other components of the proposed system as a feed forward neural network (Cho *et al.*, 2014).

2.4. BLEU

BLEU is one of the most used metrics for performance evaluation of machine translators (Cho *et al.*, 2014). Bilingual Evaluation Understudy (BLEU) is a technique for assessing a machine translation model that computes the score by comparing the model's output phrases as the candidate and the human translated words as the reference. The BLEU score ranges from 0 to 1.

3. Related Works

Chen et al. (2018) proposed a novel character-aware encoder-decoder structure. This structure had an increased encoder to produce better source word portrayals and expanded decoder with a second consideration to process source side characters in order to create better translations.

Sutskever et al. (2014a) presented encoder decoder system with LTSM neural network that initially processes the source sentence by the encoder and acquires a fixed vector representation. The decoder part then produces outputs at each time step by processing this vector representation. Activity stops when an extraordinary token demonstrating the part of the agreement is delivered

Johnson et al. (2017) developed a method for translating several languages using a single neural machine translation by inserting a token at the beginning of the source phrase and translating it to another sentence. Because of common word component language, it is feasible to interpret diverse languages using a single model. They also exhibited the zero shot problem.

Revanuru et al. (2017) created a system by applying NMT technique which have multiple models and applied this model on six Indian languages using eight different variation to train their models. As their model has a simper structure, it is simple to train.

Koehn and Knowles (2017) in this paper illustrated six challenges to NMT by comparing differences between NMT (Neural Machine Translation) and SMT (Statistical Machine Translation) after working over English-Spanish and German-English language pairs and utilizing Nematus and Moses toolkit. They tried to show that NMT still faces various issues among which domain mismatch is the most remarkable one.

Agrawal and Jain (2020) in this paper illustrated two procedures to build a device called transliteration that translates En-San language pair. The first approach used typing with keyboard and second approach used virtual keyboard. The translator uses Unicode to translate the sentences. It takes Sanskrit / Hindi as input (in English) and then using transliterate the sentence is converted into Sanskrit later it translates it in Hindi.

Bahdanau et al. (2014) demonstrated that fixed length vectors result into bottleneck which prompts a performance misfortune and they enabled a framework to naturally look through a lot of source words, thus expanding the essential encoder. Also, their model yields better outcomes on longer sentences as it doesn't depend on consolidating entire sentence to a fixed length vector.

Sennrich et al. (2015) presented that the open vocabulary issue faced by NMT can be minimised by using sub words where a variation about byte per encoding for word segmentation was introduced that might be fit about encoding open vocabularies for a traditionalist picture vocabulary for variable length sub word units.

Ha et al. (2017) proposed introduction of an intermediate pivot language to address Zero shot translation where source language is translated to pivot language and the parallel data can be then converted to source language and developed a single system with many-many MLNMT concept using only monolingual data.

Shah and Bakarola (2019) proposed using attention mechanism to improve the accuracy of NMT Seq-Seq model when producing target languages in case of working with Indian Languages. On training corpus, they have got BLEU score of 59.63 and 40.33 while on test corpus

After analyzing the performance of attention mechanism based NMT on languages such as Hindi-Bengali, Das et al. (2016) presented a Hindi-Bengali transliterator to address the named entity issue and developed post processing rules to overcome the untranslated words problem.

Verma et al. (2019) proposed using Supervised learning with attention mechanism to improve translating text from Hindi-English using NMT by using recurrent neural network mapping the input sentences to a fixed vector length sentence

Choudhary et al. (2018) proposed using pretrained Byte-Pair-Encoding and Multi BPE encoding along with multi head self-attention mechanism to overcome OOV vocabulary problem while translating to Indian languages.

Choudhary et al. (2018) proposed using word embedding along with BPE to overcome OOV problem while translating low resourced language pair like English-Tamil.

Chandola and Mahalanobis (1994) presented a tool working as automatic language translator that is based on neural networks. For Hind-English translation neural networks is employed to find out some set of ordered rules. that are often effective to other Indo-European languages if the dictionary is modified.

Narayan et al. (2014) developed an MT system based on a quantum neural network-based technique that learns the pattern of a parallel corpus by leveraging part of speech information from each word in the corpus. Narayan et al. (2014) proposed NMT system using LSTM network and bidirectional neural network in encoder model adding attention mechanism to solve lengthy sentence translation problem.

Table 1: Key points of above-mentioned papers.

No.	Author	Conclusion	Performance Evaluation
1.	Chen *et al.* (2018)	Likelihood connection tested only on few non-India language pairs.	Their outcomes demonstrate that their methodology accomplishes generous enhancements over them (up to +4.32 BLEU).
2.	Sutskever *et al.* (2014b)	Source sentence must be encoded into a fixed length vector measured which results in bottleneck.	They got 37.0 BLEU score, which is more noteworthy than the 35.8 revealed by statmt.org\matrix.
3.	Johnson *et al.* (2017)	This model works poorly on Indian Languages.	Their multilingual NMT models improve the translation nature of low-resource languages.
4.	Revanuru *et al.* (2017)	They utilized shallow system which just works with low assets. Greater information can yield awful outcomes with this system.	Their system has beated Google translation by a good margin. The BLEU score acquired are 46.47(PU- HI)35.69(GUJ-HI) 22.47(U-HI) 7.56(TA-HI).
5.	Koehn and Knowles (2017)	The result could have been much better if they used more layers or LSTM with bidirectional instead of a simple tool.	They demonstrated that, machine translation still needs to overcome out-of domain and low resource problem.
6.	Agrawal and Jain (2020)	Any word that has more than one conceivable translation, translator can't realize which significance is required.	100% precision of the framework is accomplished.
7.	Bahdanau *et al.* (2014)	Results degrade when a sentence has multiple rare words in it. Current state of art is also not considered.	Their investigation uncovered that the proposed RNN search beats the regular encoder–decoder model.
8.	Sennrich *et al.* (2015)	The proposed NMT framework for solving out of vocabulary issue just	For English→German, WDict produces highestactness(60.6%) and few OOVs (26.5%review), For English→Russian WDict pattern

No.	Author	Conclusion	Performance Evaluation
		performs useful for non-Indian dialects.	performs ineffectively for OOVs (9.2% precision;5.2% review).
9.	Ha *et al.* (2017)	This method still requires a strong cross lingual signal. It also somehow requires a parallel corpus.	Ontst2010,German→Dutch BLEU scores increased by 2.20 in case of word feature of language with similar test compared to Zero 6L (3b versus 3).
10.	Shah and Bakarola (2019)	They used Attention mechanism to deal with rare word problem.	On training corpus, they have got BLEU score of 59.63 and while 40.33 on test corpus from English-Gujarati language.
11.	Das *et al.* (2016)	They implemented post processing heuristic and Hindi-Bengali transliterator.	Their model outperforms MOSES in BLEU score: 20.41 in attention based NMT and 14.36 in MOSES.
12.	Verma *et al.* (2019)	This model has been trained with limited sentences.	Their model showed NMT is better way to translate languages.
13.	Choudhary *et al.* (2020)	OOV issue has not minimized completely.	Their model got BLEU score of 24.34 and 9.78 where Google translator got 9.40 and 5.94.
14.	Choudhary *et al.* (2018)	OOV issue has not minimized completely.	Their model outperformed google translator by 4.58 BLEU score
15.	Chandola and Mahalanobis (1994)	They only presented the result for one pair of language.	Their model did good with Indian language pair.
16.	Narayan *et al.* (2014)	The testing sentences are only a few.	The BLEU score for the system is 0.7502.The accuracy is slightly higher than google translation.

4. Methodology

4.1. Implementations of neural machine translator

We have gone through a good number of research work done regarding machine translation for Indian languages and we have observed that machine translation for North-Eastern languages are very few. As a result, we developed three neural machine translation models for the English-Assamese language pair, as we have not found any machine translation works on this language pair. Our models follow Encoder-Decoder architecture, using GRUs in the encoder and decoder and also use attention mechanism proposed by Bahdanau et al. (2014). We have used Tensorflow and Keras as frameworks in order to build our models. In order to get word embeddings of the vocabularies, we have used the Skip-gram algorithm from Word2Vec.

4.2. Dataset

We have been compiling our bilingual corpus by hand, and by making use of this collection of 1100 sentences, we have been able to successfully validate our models. The vocabulary found for the English language in this corpus is 916 words, whereas the vocabulary discovered for the Assamese language is 1429 words. After training the model with 95 percent of the data in the corpus, we tested it with the remaining 5% of the corpus.

The model and corpus statistics are as follows Table 2 and Table 3:

Table 2: Model statistics.

Type of RNN cell used	Gated Recurrent Units(GRUs)
Number of learning units each layer	1024
Learning Rate	0.0001
Type of Optimizer used	Adam
Word vector dimension	512

Table 3. Corpus statistics.

Number of parallel source sentences	2,000
Number of parallel target sentences	2,000
Source vocabulary size(words)	1,889
Target vocabulary size(words)	1,941
Training data (%)	90
Testing data (%)	10

5. Results

We have implemented our models using the above corpus mentioned, the performance statistics of the three models in Figure 2 and Table 4:

Table 4: Performance evaluation of models.

Model No.	Number of layers	Encoder	Decoder	BLEU score
1.	1	Unidirectional GRU	Unidirectional GRU	28.0271%
2.	2	Bidirectional GRU	Unidirectional GRU	34.1638%
3.	2	Unidirectional GRU	Unidirectional GRU	22.7961%

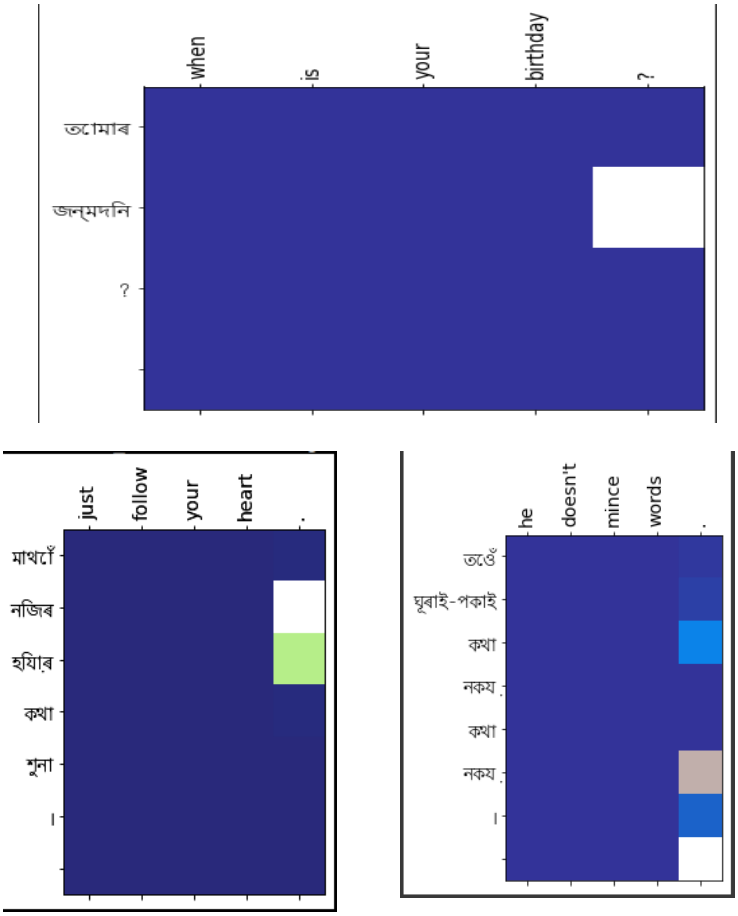


Fig. 2: Attention Visualization for the model number 1, 2 and 3 respectively.

6. Conclusion

Because of their superior performance in terms of prediction, the deep layered neural networks that we installed were selected as our solution of choice. Because they travel through a greater number of layers, each of which contains a large number of learning units, their predictive capacity is greater than that of neural networks with a single layer. This is because each layer has a large number of learning units. However, this comes at a cost, and that cost is the complexity inherent in the process of training such deep neural networks. These models have difficulties converging, which implies that if the model does not converge or if the loss percentage each epoch does not continue to drop, the model will be unable to produce accurate predictions for data that has yet to be seen. If the model fails to converge, it will be unable to generate accurate predictions about data that has yet to be seen. The model with a bidirectional GRU in the encoder and a unidirectional GRU in the decoder has been proved to be the best of the three remaining models. As a consequence of the testing, which included 55 previously unseen phrases, this model was able to attain a BLEU score of roughly 34%. This model is one of the other four models that are tabulated higher up on this page. The model that includes an attention mechanism was shown to be the most successful out of these three different models.

In this paper, we have discussed different approaches to neural machine translation architectures, their shortcomings and advantages. We have also implemented a neural machine translator. Due to minimal availability of dataset we have prepared our own dataset. We have observed that, the subword tokenization techniques are useful in dealing with Out Of Vocabulary (OOV) and rare words, and the recently developed architecture, Transformers which are found to be producing quality translation outputs, supporting more parallelization and taking less time to train. We also investigated several types of attention systems that may exist. One key finding from our investigation is that there are still few neural machine translation systems specialized to Indian languages, particularly those spoken in India's north-eastern area. We must handle the issue of Out Of Vocabulary (OOV) and unusual words in a future project, which may be partly overcome using sub-word tokenization approaches like the Byte-pair encoding technique. We want to experiment with additional neural networks, such as feed-forward neural networks, in addition to RNNs while developing NMT models.

References

Agrawal, P. & Jain, L. (2020), Anuvaadika: Implementation of sanskrit to hindi translation tool using rule-based approach, *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science),* 13(6), 1136-1151.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. DOI:10.48550/arXiv.1409.0473.

Chandola, A. & Mahalanobis, A. (1994), Ordered rules for full sentence translation: a neural network realization and a case study for Hindi and English, *Pattern recognition,* 27(4), 515-521.

Chen, H., Huang, S., Chiang, D., Dai, X., & Chen, J. (2018). Combining character and word information in neural machine translation using a multi-level attention. Paper presented at the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*. DOI:10.48550/arXiv.1409.1259.

Choudhary, H., Pathak, A. K., Saha, R. R., & Kumaraguru, P. (2018). Neural machine translation for English-Tamil. Paper presented at the Proceedings of the third conference on machine translation: shared task papers, Brussels, Belgium. DOI:10.18653/v1/W18-6459.

Choudhary, H., Rao, S., & Rohilla, R. (2020). Neural Machine Translation for Low-Resourced Indian Languages. *arXiv preprint arXiv:2004.13819*.

Christalin, N. S., Tapan, K. M., & Prakash, G. L. (2022), A novel optimized LSTM networks for traffic prediction in VANET. *Journal of System and Management Sciences*, 12(1), 461-479.DOI:10.33168/JSMS.2022.0130.

Das, A., Yerra, P., Kumar, K., & Sarkar, S. (2016). A study of attention-based neural machine translation model on Indian languages. Paper presented at the Proceedings of the 6th workshop on South and Southeast Asian natural language processing (WSSANLP2016).

Ha, T.-L., Niehues, J., & Waibel, A. (2017). Effective strategies in zero-shot neural machine translation. *arXiv preprint arXiv:1711.07893*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *arXiv preprint arXiv:1409.0473*. DOI:10.48550/ARXIV.1611.04558.

Khoa, B. T., Son, P. T., & Huynh, T. T. (2021), The relationship between the rate of return and risk in fama-french five-factor model: A machine learning algorithms approach, *Journal of System and Management Sciences,* 11(4), 47-64. DOI:10.33168/JSMS.2021.0403.

Kim, J.-Y., & Lim, C.-K. (2022), The use of sentiment analysis and latent dirichlet allocation topic-modeling (LDA) on web novel content quality factor. *Journal of System and Management Sciences,* 12(2), 236-251. DOI:10.33168/JSMS.2022.0211.

Koehn, P. & Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*. DOI:10.48550/arXiv.1706.03872.

Narayan, R., Singh, V. P., & Chakraverty, S. (2014), Quantum neural network based machine translator for Hindi to English. *Scientific World Journal,* 485737. DOI:10.1155/2014/485737.

Revanuru, K., Turlapaty, K., & Rao, S. (2017). Neural machine translation of indian languages. Paper presented at the Proceedings of the 10th annual ACM India compute conference.

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*. DOI:10.48550/arxiv.1508.07909.

Shah, P. & Bakarola, V. (2019). Neural machine translation system of Indic language-Attention based approach. Paper presented at the 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014a). Sequence to sequence learning with neural networks. *Advances in neural information processing systems,* 27(1-9).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014b). Sequence to sequence learning with neural networks. *Advances in neural information processing systems,* 27.

Verma, C., Singh, A., Seal, S., Singh, V., & Mathur, I. (2019). Hindi-English neural machine translation using attention model. *International Journal of Scientific & Technology Research,* 8(11), 2710-2714.