MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

Importance of different Machine Translation Systems for Indian Languages using Deep Learning Methodology

Ritika Goyal

Department of Computer Science & IT
Kathua Campus, University of Jammu, India
goyal.riti@gmail.com

Abstract: This paper describes the importance of different machine translation systems and status of languages used in a multilingual country like India. Machine Translation (MT) is a field of research that has been around since the birth of computers. Processing any translation whether machine or human translation, the description of text in the source language must be completely changed or transformed into the target language. In India, there is a considerable need for translating text from one language to another. There are many approaches for translating the text from one language into another. A lot of research projects are going on machine translation systems using Deep Learning Techniques between English and Indian Languages. Neural Machine Translation (NMT) is a new technique for machine translation that has led to remarkable improvements compared to rule-based and statistical machine translation (SMT) techniques, by overcoming many of the weaknesses in the conventional techniques. MT is the most difficult research work in the field Natural Language Processing (NLP) in the whole world. In the period of Artificial Intelligence, computers with ability of Deep Learning will make translation with better quality and high efficiency.

Keywords: Machine Translation, Deep Learning, NLP, Artificial Intelligence

1. Introduction

Machine Translation is the process in which computer software is used to convert a text from one language to another [1] without any human intervention. Due to globalisation of information, MT is in great demand nowadays. It is the biggest and important applications of the natural language processing (NLP). It is an instrument for reducing the language barriers. India is a multilingual country with twenty-two constitutionally recognized languages and over 2000 dialect [7]. Communication and information exchange between people is necessary not only for business purposes but also for the people to share feelings, thoughts, opinions and facts. As it is not feasible to have human translators everywhere, we need effective approaches which do this job with as little human effort as possible.

Work on MT is going on since birth of computers and is progressing very rapidly since 1990's due to availability of bilingual and multilingual corpora of languages. Since then, different approaches have been proposed for MT; Rule based translation, Knowledge based translation, Corpus based translation and hybrid translation [10]. These approaches are discussed in next section.

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

1.1 **Approaches to Machine Translation**

There are many approaches of Machine Translation system. At present, the most often used approaches of MT system are [2][3][4][7]:

1. Rule Based Machine Translation (RBMT)
2. Corpus Based Machine Translation
3. Hybrid Machine Translation
4. Deep Learning Machine Translation System

Rule Based Machine Translation: It is a knowledge- based machine translation that restores rules from bilingualism machine- understandable dictionaries and grammars based on semantic information about source and target languages. The target language is generated by the use of these rules. To implement this machine translation, extensive knowledge of both the source and target language is required. As the rules increases, the system becomes more complex and slower to translate. Preparing and producing large quantity of rules is very difficult and requires years of effort and language analysis.

Corpus Based Machine Translation: These systems use large corpora for developing Machine Translation systems. The accuracy of these systems highly depends on the quality and quantity of corpus used. The main advantage is that they can learn the translations of terminology from previous translated text. Knowledge based MT, Example Based MT and Statistical MT are the types of Corpus-Based Machine Translation Methods.

Knowledge based MT: In this approach, target text is produced and scored by a statistical model and the variables of this model are grabbed from a parallel and monolingual large corpus. The decoder will make use of these models to make translation [8]. The idea is that any sentence in a language may be a translation of a sentence in another language, but the probability changes. The purpose of this translation is to find the maximum possible sentences [9]. It does not require to be explained by people and can do translations directly in an unsupervised way of learning from the training details. Soon with the rapid increase in computer networks and easy access of data, SMT is presently the most accepted model. English-Vietnamese machine translation system is an example of KBMT.

Example Based MT: It is also known as memory-based Machine Translation. In this approach, a collection of sentences from source language is provided and produces translations of target language using point-to-point mapping [10]. It is based on the idea of reusing examples of previous translations. A database of previously analysed text is stored in the Translation Memory. The main advantage of this approach is that the results will be accurate because the texts have been retrieved from databases of actual translations produced by professional

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

translators and this approach works well for small amount of data. English to Arabic MT system is an Example based Machine Translation System.

Statistical MT: Warren Weaver in 1949, established the idea of Statistical Machine Translation (SMT). In this approach, Statistical methods are applied to generate text in target language using bilingual corpora. SMT consists of Translation Model, Language Model and Decoder. The concept of this approach came from information theory [24]. Also, this tool is inexpensive and fast as compared to Rule based tools. Another advantage is better use of resources which gives more natural transactions. N-gram based SMT is one of the examples of SMT system.

Hybrid Machine Translation: It is a method of Machine Translation in which multiple translation approaches are used within a single translation system. Neither example based nor the statistical approaches have come out to be better than Rule based approaches, though each has produced better results in some cases. For this reason, Hybrid approaches are becoming increasingly popular. It uses Rule Based Machine Translation as a base and processes the various rules with the help of statistical models. They have been successful in increasing the accuracy of the translations. The problem of word sense disambiguation [11] is also solved using Hybrid approach. TransEasy is a hybrid approach-based machine translation system. It uses example as well as statistical based approaches. Bengali to Hindi MT System has also used Hybrid approach for machine translation.

Deep Learning Machine Translation System: At present, Deep Learning plays an important role in machine learning. It is the most robust MT technology that is used to solve the complicated problems of image processing, speech, natural language processing etc. Machine translation translates the text from one language to another but with the help of deep learning, Neural Machine Translation has become the most powerful algorithm for the translation [12]. NMT uses large neural networks that uses artificial networks in a way that is based on human brain. These systems use large datasets of translated sentences to train a model which is essential in translating the text from source to target language [13]. The choice of approach merely depends on the languages being taken and the availability of computational resources. Google Translate is an example of Neural Machine Translation which is currently used in image applications, speech recognition, big data analysis etc.

2. Deep Learning and Types of Neural Networks

The availability of large volume of data, optimization algorithms for neural networks and increase in computational capacity have introduced the concept of Deep Learning in Machine Translation. With the development of Deep Learning, a new technique called Neural Machine Translation (NMT) has evolved which led to amazing improvements [15] as compared to earlier machine translation techniques [16]. Google with its Google Neural Machine Translation (GNMT) is a neural network which is available in eight languages.

Neural networks can be defined as a computing system empirical approach to machine Translation that uses large artificial neural network (ANN) to know the occurrence of large sequence of words and generate the sentences in a single integrated model.

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

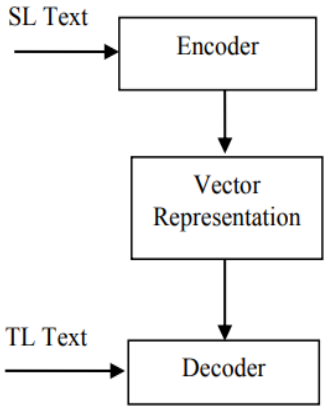

Fig.2 Neural Machine Translation

Unlike traditional translation systems, NMT is a better choice for more precise results and better performance. Deep Neural Networks is an extension of neural networks in which multiple neural network layers are processed [24] instead of one. The neural network is divided into three layers: Input layer, Middle layer (Hidden) and the outer layer. Each layer is made up of neurons (millions in number) [27] which is the elementary unit of network and which can calculate values from inputs.

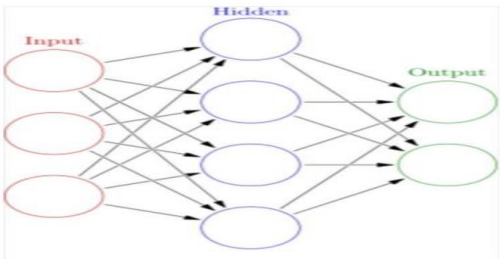

Fig 3. Simple Neural Network

2.1 Types of Neural Networks

The neural networks are classified as 20][21][22]:

Feed Forward Neural Network (FNNs): It is a basic neural network which contains several layers of processing units in which the flow control comes from input layer via hidden nodes and moves towards output layer in a feedthrough manner. The movement of data occurs only in one direction. These networks are classified as single and multilayer perceptron. The single network has a function that maps the input to an output value and the multi-layer consists of various connected layers in a directed graph. These networks are used in facial recognition algorithms.

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

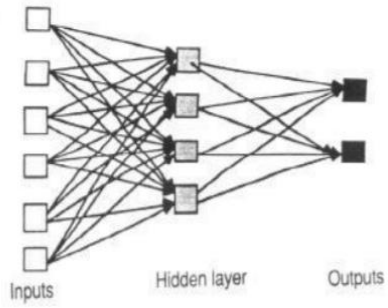

Fig. 2a) Feed Forward Neural Network

Most commonly used FNNs is **Convolutional Neural Networks (CNNs).** These networks are made up of neurons where each layer is connected to all neurons in the layer. The process of pre-processing in these networks is not much as compared to other classification algorithms. In these types of networks, the connectivity design is very much close to the connectivity network of neurons in the human brain. They are very effective in areas like image and video recognition, classification and natural language processing.

1. **Recurrent Neural Networks (RNNs)**: In earlier networks inputs and outputs are independent of each other but in RNN, they are dependent as the output of a given neuron is fed back as an input to the same node. It has a cyclic structure through which repeated sequences can be easily learned as compared to other networks. RNN has an internal memory (Hidden state) which remembers all the information which is to be calculated and through this memory output can be easily predicted.

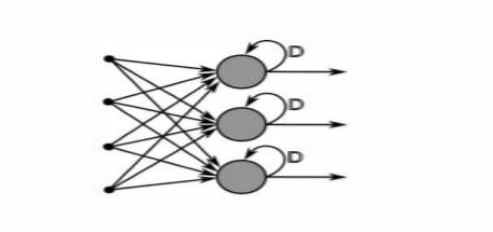

Fig. 2b) Recurrent Neural Network

Some types of RNN are:

- Long-Short Term Memory (LSTM)
- Bi-Directional RNN
-

2.1.1 **Long-Short Term Memory (LSTM)**

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

A variant of RNN called Long Short Term Memory (LSTM) [18] is known to learn problems with long range temporal dependencies, so RNNs are sometimes replaced with LSTMs in MT networks, because they may succeed better in this setting. In Figure 1, the model reads an input sentence "AB" and produces "WXY" as the output sentence. The model stops making predictions after outputting the end-of-sentence (EOS) token.

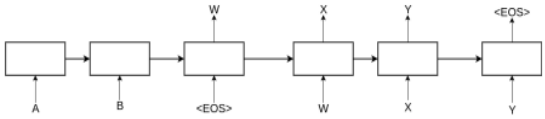

Fig.2c) LSTM

2.1.2 **Bidirectional RNN**

A bidirectional encoder is based on the idea [21] that the output at any time instant may not only depend on past data but also on future data. For example, if we are given a task to predict a missing word in a sentence, we read what is written both to the left and to the right of the missing word, to get the context. Using this idea, the LSTM is tweaked to connect two hidden layers of opposite directions to the same output. This tweaked version of LSTM is called a Bidirectional LSTM (Bi-LSTM). Bi-LSTMs were introduced to increase the amount of input information available to the network. Unlike LSTMs, Bi-LSTMs have access to the future input from the current state without need for time delays. Bi-LSTM is especially useful when the context of the input is needed. Bi-LSTMs are also used as bidirectional encoders.

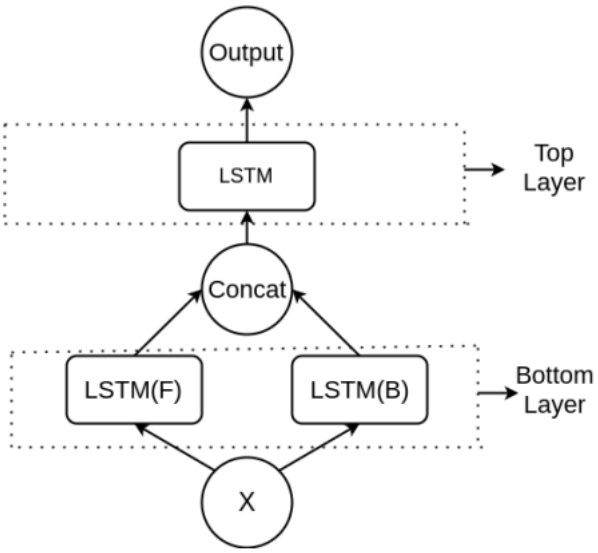

Figure 2d) Bidirectional RNN

Figure 2 d) details the use of Bi-LSTMs. Bi-LSTMs are used only in the bottom layer. The encoder in the bottom layer consists of two independent encoders, the one labelled LSTM(F)

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

encoding a normal sequence, and the LSTM(B) encoding the sequence in the reverse direction. The outputs from both these layers are concatenated and sent to the next layer labelled LSTM.

Encoder-Decoder: This type of architecture is like autoencoders which predict their own input. An encoder transforms a sentence into vector form which represents the meaning of that sentence. Firstly, word representations of both source and target language words are obtained which is fed into encoder-decoder network. The encoder network transforms these word representations into sentence embeddings. This task is done by Bi-Directional RNN which contains two RNNs [23] for calculating left-side and right-side hidden state sequences. Decoder is similar to encoder. It is responsible for predicting the words of target language by using sentence embeddings produced in encoder. The decoder generates target words based on hidden state information obtained from the encoder. When the sentences are long, it is not able to capture all the contexts linked with the source language then attention model is used to solve such problem. It is the fundamental technology inside Google translate service.

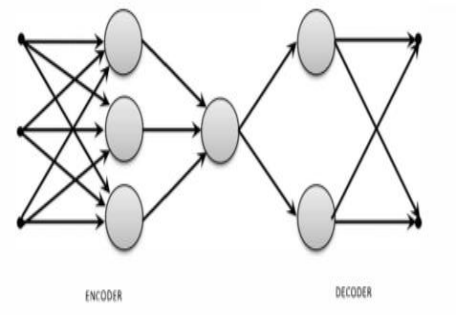

Fig. 2e) Encoder-Decoder

Let X and Y be the source and target sentence pairs respectively. The encoder RNN converts the source sentence x1,x2...xn into vectors of fixed dimensions. The decoder outputs one word at a time using conditional probability

$P(Y|X) = P(Y|X_1,X_2,X_3,...,X_m)$

Here $X_1,.....X_m$ is the equation are the fixed size vectors encoded by the encoder. Using chain rule the above equation is converted to the equation below where, while decoding, next word is predicted using symbols that are predicted till now and source sentence vectors. The above expression then becomes

$P(Y|X) = P(yi|y_0, y_1, y_2,...., y_{i-1};X_1,X_2,X_3,...,X_m)$

Each term in the distribution is represented with a softmax4 over all the words in the vocabulary.

3. Literature Review

1. Goyal and Lehal [1] developed Web based Hindi to Punjabi machine translation system. It explains the methodology followed for developing MT system for closely related languages

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

which have similar cultures and common historical roots. This system is translating any complex sentence and can be used by Newspaper agencies, any website owner of different countries or regions. The system accuracy is measured up to 95%.

2. Goyal [2] developed Hindi to Punjabi machine translation system in which Direct machine translation approach is used due to the similarity between the two languages like common root, similar alphabets, same verb patterns, almost same grammar etc. Out of the total sentences, 70.3% sentences got the score 3 i.e. intelligible, 25.1% got the score 2 i.e. generally clear and intelligible, 3.5% sentences got the score 1 i.e. they were hard to understand and 1.1% got the score 0 i.e. they were not understandable. After robust analysis, Word error rate is found to be 4.58% and sentence error rate comes to be 28.82%. The overall accuracy has been found to be 94% on the basis of intelligent test and 90.84% on the basis of accuracy test. Udupa and Faruquie [3] presented English-Hindi machine Translation system which is based on IBM models. English-Hindi parallel corpus is built and experimental results are performed which consists of 1,50,000 sentence pairs. When the maximum fertility of a word is small, the algorithms are very effective in practice.

3. Jabin, Samak and Sokphyrum [4] attempted to explain how to Translate from English to Khmer using Moses DoMY CE which is an open-asource Statistical Machine Translation tool. They used parallel data at sentence level for a translation system. Proposed system tried for only 5734 sentences for English to Khmer corpus which is saved using the same name but stored in different folders as there is no ready-made corpus available for English-Khmer. They obtained good results compared with Google Translate.

4. Singh, Kour and Jamwal [6] developed machine Translation system using Statistical approach MOSES. It requires collection of parallel text for training the system. The tools used by MOSES are GIZA++ and KenLM to build statistical language model. The parallel corpus uses 98,973 sentences and the system gives accuracy of 80% in translating English to Dogri and 87% in translating from Dogri to English. Bleu score of English-Dogri system is 22.26 and that of Dogri-English is 25.09.

5. Kadhem and Nasir [10] developed English-Arabic Example based machine Translation system in which a new approach is used to design EB by sing B+ tree. The B+ tree method is used to store the examples in EB part of the Example based Translation system to reduce redundancy of these examples and to provide efficient use of memory and to speed up the search. The lexicon of this approach is represented by using two databases. One is for storing English words and another database is used for storing the English transfer grammars. The input to the system is an English text consists of sentences. In this proposed method, 30 sentences are submitted and 24 sentences get identical instructor's translation. The accuracy scored by this system is about 80%.

Multilingual Machine Translation system for English to Urdu and Hindi using Translation Rules based approach and Artificial Neural Network was proposed by Khan and Usman [18]. The proposed technique contains ANN based training module with LM (Levenberg-Marqaurdt) algorithm which provides fast and stable convergence and used in small and medium sized problems. The evaluation score of the translated output for 500 Hindi test

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

sentences was measured using different methods like n-gram BLEU score, F-measure, meteor and precision. The system was evaluated for 500 different Hindi test sentences. The n-gram BLEU score for Hindi is 0.5903, METEOR score achieved is 0.7956 and F-score achieved is 0.7916. For Urdu, the BLEU score achieved is 0.6054, METEOR score is 0.8083 and F-score evaluated is 0.8250.

6. Revanuru, Turlapaty and Rao [21] developed Neural Machine Translation of Indian Languages. They applied NMT techniques for six Indian language pairs and train the models using eight different configurations and compared the performance of the system using evaluation metrics like BLEU, F-Measure and METEOR. The models are easier to train than deeper models because they are simple, need less resources sand require minimum time. They have achieved good accuracy by using a shallow network as compared to Google translate on a test dataset. The best model outperformed Google Translate by a margin of 17 BLEU points for English-Urdu, 29 BLEU scores for Punjabi-Hindi and 30 scores for Gujarati-Hindi translations.

7. Verma, Singh, Seal, Mathur [23] developed Hindi-English Neural Machine Translation Using Attention Model. It uses supervised learning algorithm in which two RNNs are used. One Recurrent Neural Network map the input sequence to a vector in a fixed dimension and second RNN decode the target sequence and showed that neural network translation is a better way to translate the text from source to target language. Neural network work on vectors. So, it compresses all the important information of the source language in encoder-decoder approach. BLEU score metric is used to evaluate the performance of the translation system with 500 sentences which are divided into 5 documents of 100 sentences each. The system used tensorflow 1.4 libraries for training the both NMT systems, one with baseline system and another with attention model. The BLEU score of NMT baseline system is 0.467728 and with attention model is 1.0 which are the better results than NMT baseline system.

4 Importance of Machine Translation Systems

Machine translation is an important application of natural language processing (NLP). There are large number of Machine Translation systems being developed in India as well as outside India in the past few years. English is recognized only by 3% of Indian population but still it always remains a link language in various fields like administration, education and business, science, technology. Therefore, machine translation has a greater significance in removing the communication gap within the region's social structure. Nowadays regional languages are dying everywhere. There are various valuable and knowledgeable material available on the web in English which must be translated into local languages so that local inhabitants can understand who does not speak or read English language well. Translation is not only required for business purposes but people can share their feelings, emotions, decisions in their own language. Most text such as official documents, reports, newspapers, magazines, dictionaries, books, letters, websites, blogs, articles are written in English language which is not able to understand by native people. So, there is a great need to develop such machine translation systems to make these documents readable.

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

5 Conclusion

This paper elaborates different approaches to machine translation and their related work in a more classified way. These techniques improve the quality of translation but to overcome the translation problems in different Machine Translation (MT) systems and to receive the accurate and better results of the translation process, Neural Machine Translation (NMT) is used. It has the capability to address many of the problems of traditional phrase-based translation systems and is able to produce better quality translations. NMT incorporates the full advantages of the MT systems that are used in the creation of NMT systems.

References

1. Vishal Goyal and Gurpreet Singh Lehal," Web Based Hindi to Punjabi Machine Translation System", Journal of Emerging Technologies in Web Intelligence, Vol.2, pp: 148-151, May 2010.

2. B. N. V. N. Raju and M. S. V. S. B. Raju, "Statistical Machine Translation System for Indian Languages," IEEE International Conference on Advanced Computing (IACC), pp:174-177, 2016.

3. Udupa U.R and Faruquie T.A.," An English-Hindi Statistical Machine Translation System", International Conference on Natural Language Processing, Vol. 3248, pp: 254-262, 2004.

4. Suraiya Jabin, Suos Samak , Kim Sokphyrum," How to Translate from English to Khmer using Moses ", International Journal of Engineering Inventions, Volume 3, pp: 71-81, 2013.

5. P. Dubey and Devanand, "Machine Translation System for Hindi-Dogri Language Pair", International Conference on Machine Intelligence and Research Advancement (ICMIRA), pp: 422-425, 2013.

6. Avinash Singh, Asmeet Kour and Shubhnandan S. Jamwal," English-Dogri Translation System using MOSES", Circulation in Computer Science, Vol. 1, pp: 45-49, 2016.

7. Vaishali M. Barkade and Prakash R. Devale," English to Sanskrit Machine Translation Semantic mapper", International Journal of Engineering Science and Technology", Vol. 2(10), pp: 5313-5318, 2010.

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

8. Li Peng," A survey of Machine Translation Methods", Indonesian Journal of Electrical Engineering, Vol. 11, pp: 7125-7130, 2013.

9. Bijimol T.K and Dr. John T. Abraham," A study of Machine Translation Methods", Indonesian Journal of Electrical Engineering, Vol.11, pp: 217-220, 2014.

10. Suhad M. Kadhem, Yasir R. Nasir," English to Arabic Example-based Machine Translation System", International Journal of Computers Communications & Control IJCCCE, Vol. 15, pp: 47-63, 2015.

11. Gurleen Kaur Sidhu and Navjot Kaur," English to Punjabi Machine Translation System using Hybrid approach of Word sense disambiguation and Machine Translation", International Journal of Computer Engineering and Technology, Vol. 28, pp: 387-398, 2013.

12. S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi and S. Jain, "Machine translation using deep learning: An overview," 2017 International Conference on Computer, Communications and Electronics (Comptelix), Jaipur, pp: 162-167, 2017.

13. Sandeep Saini and Vineet Sahula," Neural machine Translation for English to Hindi", International Conference on Information Retreival and Knowledge Management, pp: 1-6, 2018.

14. Shivkaran Singh, M. Anand Kumar and Dr. Soman K.P., "Attention based English to Punjabi neural machine translation", Journal of Intelligent and Fuzzy Systems, Vol.34, pp: 1551-1559, 2018.

15. Muskaan Singh, Ravinder Kumar and Inderveer Chana," Corpus based Machine Translation System with Deep Neural Network for Sanskrit to Hindi Translation", International Conference on Computational Intelligence and Data Science", Vol. 167, pp: 2534-2544, 2019.

16. B. Premjith, M. Anand Kumar and K.P. Soman," Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus: Special Issue on Natural language Processing", Journal of Intelligent Systems, Vol.28, pp: 387-398, 2019.

17. Muskan Singh, Ravinder Kumar and Inderveer Chana, "Improving Neural Machine Translation Using Rule-Based Machine Translation", International Conference on Smart Computing & Communications (ICSCC), pp:1-5, 2019.

18. Shahnawaz Khan and Imran Usman," A Model for English to Urdu and Hindi Machine Translation System using Translation Rules and Artificial Neural Network", The International Arab Journal of Information Technology, Vol. 16, pp: 125-131, 2019.

19. Er. Parveen Kumar, Er. Pooja Sharma," A study on Artificial Neural Networks", International Journal of Emerging Engineering Research and Technology", Vol. 2, pp: 143-148, 2014.

MULTIDISCIPLINARY JOURNAL OF CURRENT RESEARCH (2023) Vol. 01, No. 01, pp. 135-146

http://journalonline.org

20. Hong-hai phan-vu, Viet-trung tran, Van-nam nguyen, Hoang-vu dang, Phan-thuan do," Neural Machine Translation between Vietnamese and English: An Emperical study", Journal of Computer Science and Cybernetics, Vol. 35, pp: 147-166, 2019.

21. Karthik Revanuru, Kaushik Turlapaty, Shrisha Rao," Neural Machine Translation of Indian Languages ", ACM India Compute Conference, pp: 11-20, 2017.

22. Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan," Neural Machine Translation for Low Resource Languages using Bilingual Lexicon Induced from Comparable Corpora", Conference of the North American Chapter of the Association for Computational Linguistics", pp:112-119, 2018.

23. Charu Verma, Aarti Singh, Swagata Seal, Varsha Singh, Iti Mathur," Hindi-English Neural Machine Translation Using Attention Model", International Journal of Scientific & Technology Research, Vol.8, pp: 2710-2714, 2019.

24. Kunal Sachdeva, Rishabh Srivastava, Sambhav Jain, Dipti Misra Sharma,"Hindi to English Machine Translation: Using Effective Selection in Multi-Model SMT", International Conference on Language Resources and Evaluation, pp: 1807-1811, 2014.