The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

21. Adopting Deep Learning Techniques in Translating Ancient Indian Literature to Modern Languages

Dr. M.Santhanalakshmi,
Assistant Professor, Department of Computer Science,
Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women,
Chromepet, Chennai.
mslakshmi25@gmail.com
A.Esther Rani
B,Sc
Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women,
Chromepet, Chennai.
estherraniashok@gmail.com

Abstract

Indian Knowledge System is the systematic transmission of knowledge from ancient literature to modern science. Indian knowledge system aims to support and facilitate to solve the contemporary societal issues in several fields such as Holistic health, Psychology, Neuroscience, Nature, Environment & Sustainable development. This paper aims to close the gap between ancient Indian knowledge systems (IKS) and modern audiences around the world. The data are taken from books, websites, social media and datasets (such as Common Crawl and Wikipedia). Recurrent BiDirectional LSTMs and Attension Machanism have been used for the language translation. The main goal this research paper is to translate, preserve, and share the ageless wisdom of ancient India to people everywhere in the world.

Keywords: Recurrent BiDirectional LSTMs, Attension Machanism, Deep Learning,Natural Language Processing

1. Introduction

Machine Translation(ML) is the study of how to use computers to translate from one language into another. The concept of MT was first put forward by Warren Weaver in 1947 [1], just one year after the first computer, electronic numerical integrator and computer, was developed. From then on, MT has been considered to be one of the most challenging tasks in the field of natural language processing (NLP).

At the very beginning, MT systems were mainly designed for military applications. In 1954, Georgetown University, with the cooperation of the now well-known computer manufacturer International Business Machines (IBM) Corporation, completed a Russian–English MT experiment for the first time using the IBM-701 computer, demonstrating that the dream of MT had become true. MT was a hot topic for more than a decade after the 1954 demonstration, but the boom ended abruptly with the Automatic Language Processing Advisory Committee (ALPAC) report in 1966 [2]. With the availability of bilingual corpora, corpus-based methods became dominant after the 2000s. There are three corpus-based MT methods: example-based machine translation (EBMT), statistical machine translation (SMT), and neural machine translation (NMT). In the mid-1980s, EBMT was proposed to translate source texts by retrieving similar sentence pairs from the bilingual corpus [3]. In 1990, Brown et al. [4] proposed the idea of SMT, in which machines automatically learn translation knowledge from a large amount of data instead of relying on human experts to write rules.With strong progress in deep learning technology in speech, vision, and other fields, researchers began to apply deep learning technology to MT. In 2014, Bahdanau et al. [5] and Sutskever et al. [6] proposed end-to-end neural network translation models and formally used the term "neural machine translation." The basic idea of NMT is to map the source language into a dense semantic representation, and then generate the translation by using an attention mechanism.

The great improvements that have been achieved in both speech technologies and MT have led to simultaneous translation (ST) as another promising direction for MT.At the same time, Dong et al. [7] proposed a multilingual translation framework based on NMT, which is considered to be a breakthrough paper for multilingual translation in the history of NMT. Google also launched an NMT system [8], which was followed by other companies releasing their NMT systems. Thus, it only took about one year for NMT to be deployed online since it was first proposed in 2014. Since then, many major research laboratories (Google, Microsoft, Facebook, Huawei, etc.) have joined the research in this direction, and commercial products from companies such as Baidu have been serving hundreds of conferences. This renewed interest resulted in the First Workshop on Automatic Simultaneous Translation being held at ACL 2020 and a new ST track at the International Conference on Spoken Language Translation (IWSLT) 2020.

Major research gap is the handling of idiomatic expressions and cultural nuances, which are often lost in translation due to the lack of contextual understanding[9]. Another significant gap is the translation of domain specific terminology, which requires specialized knowledge and contextual awareness[10].

Machine translation can overcome linguistic obstacles, promote intercultural understanding, and facilitate international cooperation by tackling these issues. It creates fresh possibilities for cooperation, communication, and information sharing among various linguistic communities. Thus, achieving contextually appropriate translations in machine translation systems is a complex task that requires addressing various challenges. Key areas that require care include accurately translating technical or domain-specific vocabulary, managing idiomatic phrases and cultural subtleties, preserving grammatical correctness and syntactic coherence, and gathering and applying contextual information. . By addressing these challenges, machine translation can pave the way for effective cross-lingual communication and contribute to advancing global collaboration and understanding.

The paper focuses on a system for translating ancient languages into a universal language like English. So, that the valuable collection of ancient information can be transferred from one generation another generation. The following steps are followed to translate the ancient treasure to universal language:

Step1:To collect ancient text using primary data collection techniques.

Step2:To pre-process the ancient script using text cleaning and vectorization with one hot vector

encoding in order to convert the input text into tensors.

Step3:To generate a text translation and prediction model for translating the ancient script using LSTM into a current recognizable language i.e. English.

1.1 Terminology

Tokenization: This is the process of dividing the original text into individual pieces called tokens. Each token is assigned a unique id to represent it as a number.

Vectorization: The unique ids are then assigned to randomly initialized n-dimensional vectors.

Embedding: To give tokens meaning, the model must be trained on them. This allows the model to learn the meanings of words and how they relate to other words. To achieve this, the word vectors are "embedded" into an embedding space. As a result, similar words should have similar vectors after training.

The rest of paper is organized as follows: Section 2 provides literature survey on state-of-the-art research into various translation methods; Section 3 describes the Sources and Techniques , Section 4 Experiments done using LSTM, Section 5 explores the con-clusion and a description of future lines of research.

2. Literature Survey

The study done by[11] performed the translation of a sentence from one language to another which required a better understanding of the source and target languages. A hybrid approach was used for translation which combined example-based machine translation and transfer approaches that exhibit an accuracy of 75%. [12] proposed a Neural Machine Translation (NMT) system that used character-based embedding in combination in spite of word-based embedding. Convolution layers were used to replace the standard lookup-based word representations. BLEU points increased up to 3 points in the German-English translation task. [13] proposed an LSTM based language model and a Gated Recurrent Unit (GRU) language model. It used an attention mechanism similar to [14] from the machine translation. The study done by[15] proposed a system based on RNN and Encoder-Decoder for generating quatrains taking keywords as an input. The system learns the semantic meaning in the sentence and learns semantic meaning among the sentences in the poem. [16]

a machine translation technique using deep learning, Tanaka corpus was used to convert the Japanese language into the English language. In the above approach neural machine translation was used which belonged to a family of encoder–decoders in which the encoder encodes a source sentence into a fixedlength vector called tensors from which a decoder generates a translation.[17] proposed a model using RNN-LSTM for translating Arabic text into English. COCO caption dataset was built, the performance of the proposed model on the test dataset gave a result of 46.2 for the BLEU-1 score.

3. Sources and Techniques

In nature, ancient text translators are made up of different modules. It can be implemented using deep learning, that convert the source text into a target language. A Neural Network (NN) has been used for the development of the model. The model consists of an encoder and a decoder, which involves running two LSTM Recurrent Neural Networks which work together simultaneously to transform one sequence into another. An Encoder network condenses an input sequence into a vector, a Decoder network unfolds that vector into a new sequence. LSTM is a type of RNN model that computes the probability of occurrence of words in a reference text.

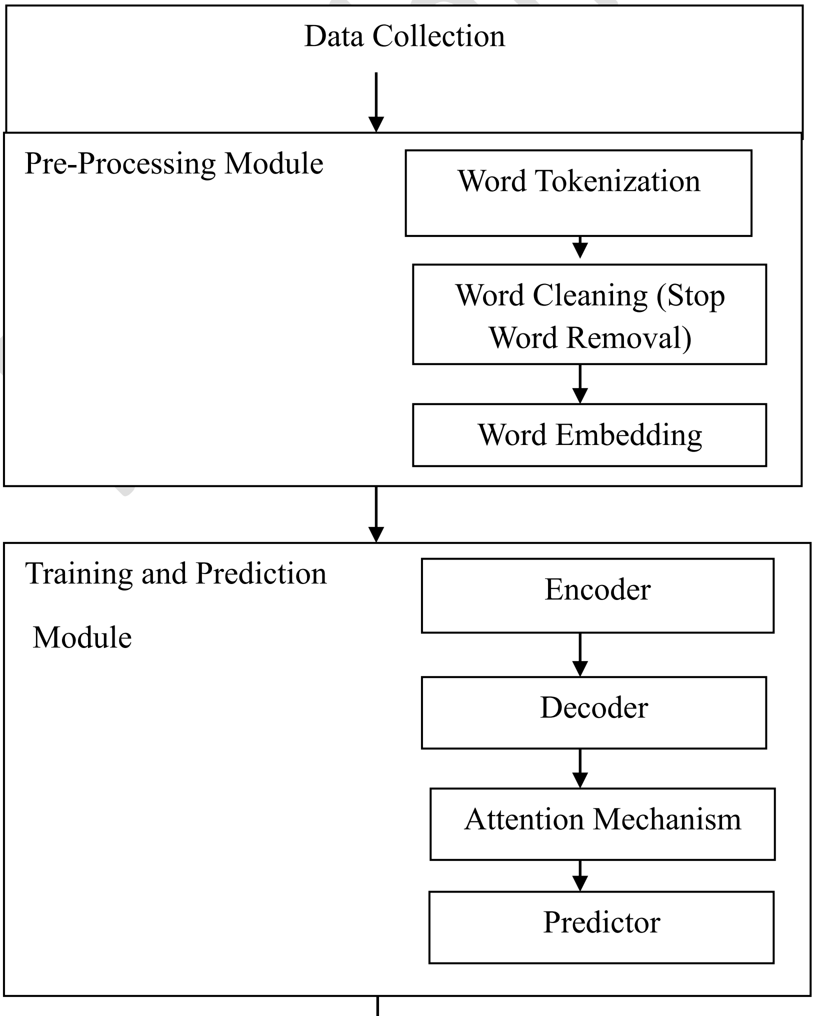

Figure 1: Conceptual diagram for language translation

3.1 Data Set

Itihāsa is a Sanskrit-English translation corpus containing 93,000 Sanskrit shlokas and their English translations extracted from M. N. Dutt's seminal works on The Rāmāyana and The Mahābhārata. The data folder contains the randomized train, development, and test sets. The original extracted data can be found in JSON format.

This work introduces Itihasa, a large-scale translation dataset containing 93,000 pairs of Sanskrit shlokas and their English translations. The shlokas are extracted from two Indian epics viz., The Ramayana and The Mahabharata. We first describe the motivation behind the curation of such a dataset and follow up with empirical analysis to bring out its nuances. We then benchmark the performance of standard translation models on this corpus and show that even state-of-the-art transformer architectures perform poorly, emphasizing the complexity of the dataset.[18]

4. Methodology

The translation and prediction model comprises vectorization and training i.e. taking source text as input and converting it into dense tensors which are further converted into target text with the help ofencoder-decoder units. Vectorization further comprises of text cleansing and one-hot encoding, whichinvolves converting the source code.

The preprocessing involves text cleansing that reduces the size of the source text by removing stop words, punctuations and finally converting all the upper case into lower case. As the text is very ancient and due to very low level of knowledge of the source language i.e., Sanskrit.

The training and prediction technique used in this study comprises of three functions, the first function converts characters to numeric indexes so that each character in the text has a unique index, whereas the second function converts those indexes into vectors. Finally, the same thing has been done for the whole pair of Sanskrit to English, so the generated output is in the form of vector using Pytorch.

After the vectors were created, the training section took the vectors as input and converted them into the target language i.e., English. It was designed using a Neural Network (NN) consisting of an encoder and a decoder, that involved running two RNN simultaneously to transform one sequence to another[19]. An encoder network condenses an input sequence into a vector, and a decoder network unfolds that vector into a new sequence. The encoder of a seq2seq network is an RNN that outputs some value for every word from the input sentence. For every input word the encoder outputs vectors and a hidden state, and uses the hidden state for the next input word.

Initially, the input is given to the encoder in the form of tensors which is further passed to the embedded layer that is used to give the dense representation to the words and their complicated meaning, input to the layer is integer encoded in order to represent each word uniquely and the same is given to Gated Recurrent Unit (GRU)[20] which is a kind of LSTM that intends to solve the problem of vanishing gradients, another input to GRU[21] is from a

previously hidden layer, but as it is the first layer so the previously hidden layer tends to be empty and the result is further transferred to the output layer and hidden layer until the end of the sentence is encountered after which input is given to next layer.

The decoder takes the input from previous hidden layer and output of the encoder which is passed to the embedding layer, furthermore, the embedded output is passed to activation function Rectified Linear Unit activation function (ReLU) to avoid overfitting[22] and the output of the function is passed to softmax which calculates a probability for every possible class, thereafter, the output is provided to the hidden layer until the end of the sentence is reached after which the output is provided to the output layer.

The output of the decoder layer is transferred to the attention decoder that is used to improve the performance of the encoder-decoder translation module in parallel, it gets the input from the encoder along with the decoder and a previously hidden layer of the decoder. The input of the decoder layer is given to the embedded layer that again gives a dense representation of tensors, which is further provided to the dropout layer to avoid overfitting, afterwards, this embedded output is given to the attention layer that produces attention weights for each tensor that are further given as input to Batch Matrix Multiplication (BMM) along with embedded output that is further passed to GRU to solve the problem of gradient vanishing and keeps on giving the input to the previously hidden layer until the end of the sentence is encountered. The training and prediction module are used to predict tensors into source language, by using the mapping done by the decoder, to finally convert tensors into a text file

of the target language.

Table 1. Corpus of Sanskrit to English text

Input	तस्यां चीरं वसानायां नाथवत्यामनाथवत्। प्रचुक्रोश जनः सर्वो धिक् त्वां दशरथं त्विति ॥
Output	She was dressed in bark clothes like an orphan with a master All the people

	cried out 'Woe to you Dasaratha'
Input	एतां पश्य दुराधर्षां मायाबलसमन्विताम्। विनिवृत्तां करोम्यद्य हृतकर्णाननासिकाम्॥
Output	Look at this invincible lady endowed with the power of magic Today I shall make her withdraw her ears and nose

5. Conclusion

There are many techniques are there to translate languages. According to a limited bilingual corpus and a high number of difficult non-translated words, as well as the lack of available language knowledge, any issues had to be resolved when translating ancient texts into English. Also, low-resource source languages can be translated into universal languages with the help of the translator.

6.Annexure

Abbreviation	Nomenclature
ML	Machine Learning
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
BLEU	Bilingual Evaluation Under Study
NMT	Neural Machine Translation
CNN	Convolution Neural Network
GRU	Gated Recurrent Unit
BMM	Batch Matrix Multiplication
HTML	Hyper Text Markup Language
RNN	Recurrent Neural Network
ReLU	Recursive LU Rectified Linear Unit activation function

7. References

[1]W. Weaver Translation. Mach Transl Lang, 14 (1955), pp. 15-23

[2]J. Hutchins ALPAC: the (in) famous report S. Nirenburg, H.L. Somers, Y.A. Wilks (Eds.), Readings in machine translation, MIT Press, Cambridge (2003)

[3]M. Nagao A framework of a mechanical translation between Japanese and English by analogy principle A. Elithorn, R. Banerji (Eds.), Proceedings of the International NATO Symposium on Artificial and Human Intelligence, Elsevier North-Holland, Inc, New York City (1984), pp. 173-180

[4] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, et al. A statistical approach to machine translation Comput Linguist, 16 (2) (1990), pp. 79-85

[5] D. Bahdanau, K. Cho, Y. Bengio Neural machine translation by jointly learning to align and translate

[6]I. Sutskever, O. Vinyals, Q.V. Le Sequence to sequence learning with neural networks Proceedings of the 27th International Conference on Neural Information Processing Systems; 2014 Dec 8–13; Montreal, QC, Canada (2014) Proceedings of the 3rd International Conference on Learning Representations; 2015 May 7–9; San Diego, USA (2015)

[7] D. Dong, H. Wu, W. He, D. Yu, H. Wang Multi-task learning for multiple language translation Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015 Jul 26–31; Beijing, China (2015)

[8].Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: bridging the gap between human and machine translation. 2016. arXiv: 1609.08144.

[9]. I.Rivera-Trigueros Machine translation systems and quality assessment: a systematic review Comput. Humanit., 56 (2022), pp. 593-619

[10]A. Núñez-Marcos, O. Perez-de-Viñaspre, G. Labaka A survey on Sign Language machine translation Expert Syst. Appl., 213 (2022), Article 118993

[11]P.Y. Huang, F. Liu, S.-R. Shiang, J. Oh, and C. Dyer, "Attention-based Multimodal Neural Machine Translation.", 2018.

[12]M. Zhang, Y. Zhang, and D.-T. Vo, "Gated Neural Networks for Targeted Sentiment Analysis." [Online]. Available: www.aaai.org.

[13]G. Lin and W. Shen, "Research on convolutional neural network based on improved Relu piecewise activation function," in Procedia Computer Science, 2018, vol. 131, pp. 977–984, https://doi. org/10.1016/j.procs.2018.04.239.

[14]Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y.

(2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. https://arxiv.org/abs/1406.1078

[15]Maitra, D. sen, Bhattacharya, U., & Parui, S. K. (2015). CNN based common approach to handwritten character recognition of multiple scripts. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2015-Novem, 1021–1025. https://doi.org/10.1109/ ICDAR.2015.7333916

[16]Nurseitov.D, Bostanbekov.K, Kanatov.M, Alimova.A 1,2 , Abdallah.A, Abdimanap.G (2021). “Classification of handwritten names of cities and Handwritten text recognition using various deep learning models”, Advances in Science, Technology and Engineering Systems Journal Vol. 5.

[17]Ray, A., Rajeswar, S., & Chaudhury, S. (2015). Text recognition using deep BLSTM networks. ICAPR 2015 - 2015 8th International Conference on Advances in Pattern Recognition. https://doi. org/10.1109/ICAPR.2015.7050699

[18] Rahul Aralikatte, Miryam de Lhoneux, Anoop Kunchukuttan, and Anders Søgaard. 2021. Itihasa: A large-scale corpus for Sanskrit to English translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 191–197, Online. Association for Computational Linguistics.

[19]Purwarianti.A, Yayat.D, Fakultas.P, “Experiment on a Phrase-Based Statistical Machine Translation Using PoS Tag Information for Sundanese into Indonesian”, International Conference on Information Technology Systems and Innovation (ICITSI),p.p 1-6, 2015.

[20]Fachrurrozi, M., Yusliani, N., & Agustin, M. M. (n.d.). "Identification of Ambiguous Sentence Pattern in Indonesian Using Shift-Reduce Parsing", 2014

[21]Markou, K. et al. (2021). A Convolutional Recurrent Neural Network for the Handwritten Text Recognition of Historical Greek Manuscripts. In: , et al. Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science(), vol 12667. Springer, Cham. https://doi.org/10.1007/978-3-030-68787-8_18

[22] Gautam N. & Chai S. (2020). Translation into Pali Language from Brahmi Script. In: Sharma D.K., Balas V.E., Son L.H., Sharma R., Cengiz K. (eds) Micro-Electronics and Telecommunication Engineering. Lecture Notes in Networks and Systems, vol 106. Springer, Singapore. https://doi. org/10.1007/978-981-15-2329-8_12.