IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

San-Eng: Sanskrit to English Translator using Machine Learning

Shetty Ramakrishna Mohan, Rohan S Bhat, Ranjith V Shetty, Aniruddha
Department of Computer Science and Engineering,
Canara Engineering College, Benjanapadavu, India

Abstract: *In the past decade, machine learning has made great strides in improving automatic translation. In general, machine learning algorithms have been able to achieve better translation quality by learning from large amounts of data. One approach to machine translation is to use a statistical machine translation (SMT) system. SMT systems learn to translate by statistical methods, using a large parallel corpus of text in multiple languages. The most successful SMT systems are based on the translation model known as the phrase-based translation model. In recent years, a new approach to machine translation has emerged, known as neural machine translation (NMT). NMT systems use artificial neural networks to learn to translate. NMT systems have shown to be very successful in translating between languages that are closely related, such as English and Hindi or Reverse is also true. In recent years, a new approach to machine translation has emerged, known as neural machine translation (NMT). NMT systems use artificial neural networks to learn to translate. NMT systems have shown to be very successful in translating between languages that are closely related, such as English and Hindi or Reverse is also true.*

Keywords: Machine Learning, Statistical machine translation, Neural machine translation, Encoder-Decoder, Optical Character Recognition

I. INTRODUCTION

Sanskrit is a classical language of India with a rich literary tradition. As the language is no longer widely spoken, there is a need for accurate Sanskrit to English translation to make its literature and knowledge accessible to a wider audience. Machine learning models can be used to translate Sanskrit to English.One approach is to use Optical Character Recognition (OCR) to extract text data from Sanskrit documents or images, which can then be used as input to a translation model. OCR technology is able to recognize the Sanskrit characters and convert them into a digital format that can be read by a computer.The translation model could be an encoder-decoder neural network. The encoder would convert the Sanskrit text into a numerical representation, and the decoder would use that representation to generate the corresponding English translation.

The need for Sanskrit to English translation arises from the rich literary tradition of Sanskrit, which includes ancient texts on philosophy, religion, medicine, and science. As the language is no longer widely spoken, access to these texts is limited to those who can read Sanskrit. Translating these texts into English would make them accessible to a wider audience, helping to preserve and promote this valuable cultural heritage. Additionally, Sanskrit is still used in some religious and cultural contexts, and translating Sanskrit texts into English can help bridge cultural and linguistic barriers.

II. RELATED WORK

Machine translation model for effective translation of Hindi poetries into English

The proposed approach is a hybrid machine translation method for translating Hindi poems into English with a focus on word sense disambiguation. The method uses a trie-based dictionary to store words, phrases, and expressions in Hindi and employs syntactic and handcrafted rules to identify accurate words. The system corrects erroneous words by providing a list of valid alternatives. The proposed method performs better than existing machine translation methods. The algorithm is divided into several phases, including scanning, POS tagging, translation, and reorganization. The system utilizes a scanner to read 19th-century Hindi source material and convert it into a series of tokens. The tokens are then analyzed for word sense disambiguation, and the most likely interpretation is selected. The approach can translate complex Hindi sentences into English and can be further improved with future research. The proposed

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

approach enhances the translation process, allowing each segment to be interpreted independently and then reassembled to form the target sentence according to the translation summary.

Neural Machine Translation for Indian Language

The paper discusses the use of Neural Machine Translation (NMT) systems for bridging communication barriers between people of different linguistic backgrounds. The NMT systems have been trained, tested, and evaluated for English to Tamil, English to Hindi, and English to Punjabi translations, using parallel corpora. The results show that NMT produces fluent translations and that the performance improves with an increase in training data and length of test sentences. The effectiveness of translation largely depends on the size of the training corpus, score functions used for computing attention of each source state, and skillful selection of system parameters. The author suggests that optimization of system parameters and a better understanding of target language constructs can also help improve translation quality.

Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus

The author of this paper proposes a neural machine translation (NMT) system to translate between English and four Indian languages: Malayalam, Hindi, Tamil, and Punjabi. To model the translation system, four parallel corpora were collected from different sources and cleaned. The NMT architecture used LSTM networks and bi-directional RNNs in the encoder network, and an attention mechanism was employed to address lengthy sentences in English–Malayalam and English–Hindi corpora. The obtained models were evaluated using BLEU scores and manual metrics, showing deep neural networks improving translation quality and the system was able to perceive long-term contexts in the sentences. The author contributes to the machine translation research community with a large, diverse corpus of parallel sentences with similar meanings and covering different domains, along with linguistic features to improve translation quality. However, the length of the sentences should be appropriate, as deep learning architectures have difficulty extracting long dependencies present in long sentences. The performance of four NMT systems was tested using a separate data set containing 562 English sentences from different domains, and the author concludes that to ensure better translations, the corpus should be of large size, include parallel sentences of diverse domains and be of an appropriate length.

Sanskrit to Universal Networking Language EnConverter System based on Deep Learning and Context-Free Grammar

The authors of this paper proposed an extension of a machine translation system (MTS) for Sanskrit to UNL (Universal Networking Language) translation. The system includes a stemmer, neural network for POS (part-of-speech) tagging, Sanskrit grammar, and CYK parser, and is divided into seven layers each performing a different task. The proposed system has reported an average BLEU score and average fluency score with an overall efficiency, and is capable of resolving 46 UNL relations. The authors suggest future enhancement of the system with deep neural networks and a parallel corpus of Sanskrit sentences and UNL expressions trained on BiLSTM networks for translating Sanskrit compound sentences.

The review also discusses various machine translation systems based on UNL approach, including systems for Indian languages like Hindi, Punjabi, Tamil, and Malayalam, as well as other world languages such as English, Chinese, French, and Russian. The review also covers neural machine translation (NMT) systems, which perform end-to-end translation and use recurrent neural networks (RNNs) and encoder-decoder approach for translation

Machine Translation Systems and Quality Assessment: A Systematic Review

The paper presents a systematic literature review of Machine Translation (MT) systems for the English-Spanish language combination. The review identifies neural MT as the predominant paradigm in the current MT scenario, with Google Translator being the most used system. Most of the works use either automatic or human evaluation to assess MT, and only 22% combine both types of evaluation. Additionally, more than a half of the works include error classification and analysis, an important aspect for improving the performance of MT systems. The study was limited by the selection criteria of language pair and at least one of the authors belonging to the field of translation or similar,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

which resulted in a significantly reduced sample size. However, the study concludes that MT is a growing area with great potential for overcoming language barriers and increasing the productivity of the translation process .The paper also notes that in the current MT scenario, neural MT is better than statistical MT, as observed in the main MT evaluation forums (WNT 2015) and confirmed by the adoption of neural technologies by the main MT companies such as Google, or Microsoft. Deep learning, which is popular nowadays, has not been widely used in the studies, but its results are somewhat lower than Google's. Therefore, it would be advisable to include deep learning in similar research and compare its results with those of the current predominant system.

Optimal Word Segmentation for Neural Machine Translation into Dravidian Languages

The paper compares the effectiveness of Linguistically Motivated Vocabulary Reduction (LMVR) and Sentence Piece (SP) for sub word segmentation when translating from English to four Dravidian languages (Kannada, Malayalam, Tamil, and Telugu) using a Transformer architecture for Neural Machine Translation (NMT). The results indicate that SP is the best choice for sub word segmentation, and larger sub word vocabularies lead to higher translation quality. While BLEU scores were mixed, CHRF scores suggest that SP remains the best option for all tested language pairs. The paper also found interesting differences among the four target languages, with Kannada being the most challenging language to translate. Additionally, LMVR results in shorter sub-words for every language and dictionary size, but the widening difference between LMVR and SP with larger dictionary sizes is not easily explained. The paper concludes that there is still much room for improvement in translating Dravidian languages and invites further research.

III. LITERATURE REVIEW

Neural Machine Translation (NMT) is a popular machine learning technique used for translating texts from one language to another. Currently, tools are available to translate most popular languages like German, Hindi, Punjabi, and Spanish, yet there is still no successful tool for translating Sanskrit. Sanskrit is an important language for the Hindu religion and is still used to this day. In order to create an effective automated written translator of full sentences, San-Eng has been developed. This machine translation model, Encoder-Decoder LSTM, works by understanding each word in a sentence based on its understanding of previous words and the context of the sentence. Previous research has found two methods for translating Sanskrit to English, Statistical Machine Translation and a general algorithm of top-down processing. The aim of San-Eng is to develop an automated Sanskrit-English sentence translation model

3.1 Real-World Applications

The application of machine learning models for Sanskrit to English translation, specifically using OCR to extract data from images and then applying encoder-decoder techniques for translation, has potential use cases in a variety of real-world scenarios. One potential application could be in the digitization of ancient Sanskrit texts, which are often only available in hard copy form. By using OCR and machine translation, these texts could be made more widely available to researchers and the general public, without the need for physical access to the original documents. Another possible application could be in the field of cross-cultural communication. With increased globalization, there is a growing need for accurate and efficient translation between languages, including less widely spoken languages such as Sanskrit. By using machine learning models to automate the translation process, communication between speakers of different languages can be made more accessible and effective. Furthermore, the application of OCR and machine translation can be useful in industries such as education and tourism. For instance, Sanskrit texts could be more easily integrated into language learning materials, or visitors to historical sites in India could access translated information about the site's significance in their own language. Overall, the application of machine learning models for Sanskrit to English translation using OCR and encoder-decoder techniques has the potential to facilitate the preservation and dissemination of cultural and historical knowledge, enhance cross-cultural communication, and support the development of educational and tourism materials.

3.2 Future Research Directions

There are several future research directions for Sanskrit to English translation using machine learning models and OCR technology. Firstly, improving the accuracy of OCR for Sanskrit texts can lead to better results in the overall translation

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

process. OCR for Sanskrit poses several challenges due to its complex script, ligatures, and diacritical marks. Hence, developing OCR models specifically for Sanskrit can be an important research direction. Secondly, incorporating more linguistic knowledge and context can improve the translation quality. Sanskrit has a rich morphology, syntax, and semantics that are not fully captured by current machine learning models. Incorporating such linguistic knowledge into the translation models can lead to better translations. Thirdly, exploring the use of unsupervised machine learning techniques can be a promising research direction. Unsupervised techniques can learn from unlabeled data and can potentially reduce the reliance on annotated datasets, which are scarce for Sanskrit. Fourthly, developing a more comprehensive evaluation framework for Sanskrit to English translation can be an important research direction. Current evaluation metrics such as BLEU do not capture the quality of translation accurately for Sanskrit, which has a different syntax and semantics compared to English. Developing better evaluation metrics that consider the unique features of Sanskrit can lead to more accurate evaluation of translation models. Finally, exploring the use of transfer learning techniques can be an important research direction. Transfer learning has shown promising results in several natural language processing tasks and can potentially reduce the reliance on large annotated datasets. Fine-tuning pre-trained models on Sanskrit can lead to better translation results with limited annotated data.

IV. METHODOLOGY

OCR stands for Optical Character Recognition, which is the technology that enables machines to recognize and interpret printed or handwritten text from an image, scanned document or a physical paper. OCR systems use advanced algorithms and machine learning models to identify characters, words and sentences from an image and convert them into machine-readable text. OCR is an important technology for digitizing documents and making them searchable and editable

Encoder-decoder is a neural network architecture used in machine translation to transform an input sequence into an output sequence. The encoder component takes in the input sequence and generates a hidden representation, while the decoder component uses this hidden representation to generate the output sequence. This approach allows the model to learn the semantic representation of the input sequence, which is used to generate the output sequence. The encoder-decoder architecture has been shown to be highly effective in machine translation tasks.

4.1 PREPROCESSING

A. Preprocessing of the Dataset

- Data has been preprocessed to remove punctuations, digits and extra spaces.
- Unique words have been processed into a dictionary with an index value allotted to each unique word
- All the sentences have been converted to lower case.

4.2 Text Extraction

- Python's tesseract library/utility has been used for the purpose
- Tesseract has been configured to read Hindi texts and returns the text it encountered.

4.3 Text Translation

- One hot encoding has been used to represent words as vectors
- Encoder encodes the given text into one hot encoded vectors and the decoder decodes it back to the original word.
- This follows word to word translation.
- In case of image, text will be extracted from image using pytesseract OCR and the extracted text will be fed to the model
- There is also an option to give text as inputs directly
- Model in this way is an encoder-decoder architecture and will produce appropriate translations as outputs

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

A. Collection of Data:

Data source: https://www.kaggle.com/code/aiswaryaramachandran/english-to-hindi-neural-machine-translation/data

- This dataset contains Hindi sentences and their English translation
- Total count: 127608 sentences with translations

B. Preparing

Datasource:
https://www.kaggle.com/code/rmuhammed97/denoise-images-using-autoencoders
Contains: Train images: 144, Test images: 72,Cleaned trained images (validation): 144

C. Blur Detection Dataset

- Images used for this purpose have been scraped through web using Selenium
- Around 200 Sharp images have been scraped initially
- Only around 30% of them were useful
- They have been further augmented to generate more of image data
- All the sharp images have been blurred using cv2.blur()
- Total dataset size post augmentation:
  Blur images: 111
  Sharp images: 111

Algorithm:

Encoder-Decoder algorithm is used to translate Sanskrittext to English.

Evaluation with test set:

Several image are passed through the model to check whether the used algorithm gives the correct result or not.

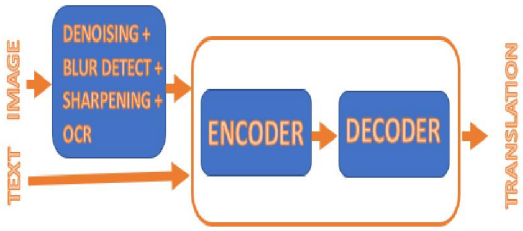

Fig.1. System Architecture Diagram

4.4 Result

Here below, we attached the screenshots of one example that we considered. Finally, the analysis and classification are displayed in an easily understandable manner. The results are displayed in the form of graphs as shown below:

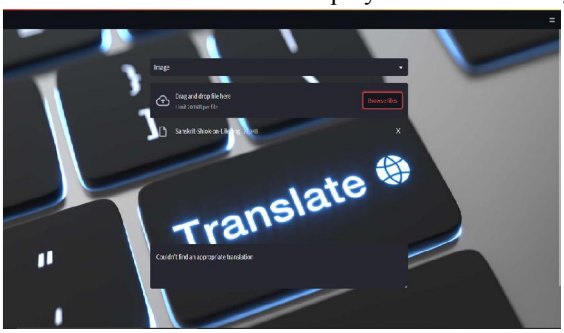

Fig.2.UI Screenshot

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

The above Screenshot shows the UI of the project where user can give input as image or Sanskrit text.

Fig.3.UI Screenshot with result

The above Screenshot shows the result of the user text input and the output

V. CONCLUSION AND FUTURE WORK

The study presented a methodology for Sanskrit to English translation using a machine learning model to extract data from images through OCR and then using an encoder-decoder technique to translate it to English. The proposed method includes the use of Tesseract OCR for data extraction, preprocessing techniques for improving image quality, and an encoder-decoder architecture with attention mechanism for translation. Future work can involve the development of an improved OCR engine for better text extraction from images, incorporating more advanced techniques for image pre-processing to further improve the quality of the images and therefore the accuracy of text extraction. Additionally, the training of the model can be improved by incorporating more data and exploring different neural network architectures. Another area of future work is the extension of the proposed method to other languages to improve the accuracy and efficiency of machine translation

REFERENCES

[1]. S. Kondo, "Machine Translation with Pre-specified Target-side Words Using a Semi-autoregressive Model," no. 2019, pp. 68–73, 2021.

[2]. R. K. Chakrawarti, J. Bansal, and P. Bansal, "Machine translation model for effective translation of Hindi poetries into English," J. Exp. Theor. Artif. Intell, vol. 34, no. 1, pp. 95–109, 2022, doi: 10.1080/0952813X.2020.1836033.

[3]. A. Pathak and P. Pakray, "Neural machine translation for Indian languages," J. Intell. Syst., vol. 28, no. 3, pp. 465–477, 2019, doi: 10.1515/jisys-2018-0065.

[4]. B. Premjith, M. A. Kumar, and K. P. Soman, "Neural machine translation system for English to Indian language translation using MTIL parallel corpus," J. Intell. Syst., vol. 28, no. 3, pp. 387–398, 2019, doi: 10.1515/jisys-2019-2510.

[5]. Sitender and S. Bawa, "Sanskrit to universal networking language EnConverter system based on deep learning and context-free grammar," Multimed. Syst., 2020, doi: 10.1007/s00530-020-00692-3.

[6]. Rivera-Trigueros, "Machine translation systems and quality assessment: a systematic review," Lang. Resour. Eval., vol. 56, no. 2, pp. 593–619, 2022, doi: 10.1007/s10579-021-09537-5

[7]. O. Hellwig, S. Sellmer, and S. Nehrdich, "Obtaining more expressive corpus distributions for standardized ancient languages," CEUR Workshop Proc., vol. 2989, pp. 92–107, 2021. A.

[8]. O. Hellwig, S. Scarlata, E. Ackermann, and P. Widmer, "The treebank of Vedic Sanskrit," Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., pp. 5137– 5146, 2020.

[9]. O. Hellwig, "Dating and Stratifying a Historical Corpus with a {B}ayesian Mixture Model," Proc. LT4HALA 2020 - 1st Work. Lang. Technol. Hist. Anc. Lang., no. May, pp. 1–9, 2020, [Online]. Available: https://aclanthology.org/2020.lt4hala-1.1

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

[10]. O. Hellwig and S. Sellmer, "Detecting Diachronic Syntactic Developments inPresence of Bias Terms," Proc. Second Work. Lang. Technol. Hist. Anc. Lang., no. June, pp. 10–19, 2022, [Online]. Available: https://aclanthology.org/2022.lt4hala1.2

[11]. P. Dhar, A. Bisazza, and G. Van Noord, "Optimal Word Segmentation for," no. 2012, pp. 181–190, 2021.

[12]. M. Singh, R. Kumar, and I. Chana, "Machine Translation Systems for Indian Languages: Review of Modelling Techniques, Challenges, Open Issues and Future Research Directions," Arch. Comput. Methods Eng., vol. 28, no. 4, pp. 2165–2193, 2021, doi: 10.1007/s11831-020-09449-7.

[13]. O. Hellwig, "Dating Sanskrit texts using linguistic features and neural networks," IndogermanischeForschungen, vol. 124, no. 1, pp. 1–46, 2019, doi: 10.1515/if2019-0001.

[14]. R. Haque, M. Hasanuzzaman, and A. Way, "Investigating terminology translation in statistical and neural machine translation: A case study on English-to-Hindi and Hindi-to-English," Int. Conf. Recent Adv. Nat. Lang. Process. RANLP, vol. 2019- Septe, no. 2017, pp. 437–446, 2019, doi: 10.26615/978-954-452-056-4_052

[15]. S. R. Laskar, A. F. U. R. Khilji, D. Kaushik, P. Pakray, and S. Bandyopadhyay, "Improved English to Hindi Multimodal Neural Machine Translation," WAT 2021 - 8th Work. Asian Transl. Proc. Work., pp. 155–160, 2021, doi: 10.18653/v1/2021.wat-1.17.

[16]. O. Hellwig and S. Nehrdich, "Sanskrit word segmentation using character-level recurrent and convolutional neural networks," Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, pp. 2754–2763, 2018, doi: 10.18653/v1/d18- 1295.

[17]. W. Lu, "Word sense disambiguation based on dependency constraint knowledge," Cluster Comput., vol. 22, pp. 7549–7557, 2019, doi: 10.1007/s10586-018-1899-3.

[18]. A. R. Pal, D. Saha, N. S. Dash, and A. Pal, "Word Sense Disambiguation in Bangla Language Using Supervised Methodology with Necessary Modifications," J. Inst. Eng. Ser. B, vol. 99, no. 5, pp. 519–526, 2018, doi: 10.1007/s40031-018-0337-5.

[19]. S. R. Laskar, A. F. U. R. Khilji, D. Kaushik, P. Pakray, and S. Bandyopadhyay, "Multimodal Neural Machine Translation for English to Hind," WAT 2021 - 8thWork. Asian Transl. Proc. Work., pp. 109–113 2021, Available: https://aclanthology.org/2020.wat-1.11

[20]. W. Lu, "The Vedic corpus as a graph. An updated version of Bloomfield's Vedic Concordance," Cluster Comput., vol. 22, pp. 754–775, 2019, doi: 10.1009/s10556- 018-1899-3