Machine Translation for Sanskrit language

1st Semester of 2022-2023

Ilicea Anca Stefania
anca.ilicea@s.unibuc.ro

Abstract

Sanskrit language: Sanskrit is an ancient Indo-Aryan language that originated in the Indian subcontinent. It holds a significant position in Indian history, culture, and religion. This paper aims to investigate a dataset comprising translated texts from Sanskrit to English. Our objective is to delve into previous research findings related to Sanskrit-English translation and explore the feasibility of leveraging a pre-trained machine translation model for our chosen dataset. By doing so, we aim to enhance our understanding of the translation process from Sanskrit to English and potentially discover new insights or improvements in this domain.

1 Introduction

Sanskrit, recognized as one of the most ancient languages known to mankind, holds a significant position in history due to its extensive influence on a wide range of Indo-European languages. The profound impact of Sanskrit can be observed in the linguistic development and evolution of numerous languages that fall under the Indo-European language family. I found this topic interesting to approach because it has a highly systematic and regular grammar with precise rules for word formation, syntax, and phonetics, also because it holds a wide range of words and expressions, some of these represents specialized terms for various domains such as philosophy, mathematics, astronomy, medicine, and literature.

2 Data acquisition

The dataset used in this study, called rahular/itihasa, was obtained from Hugging Face. It comprises around 75k training data samples, 11k test data samples, and 6k validation data samples. To accommodate GPU capacity limitations on my machine, I have specifically selected 20k training samples, 4k test samples, and 1k validation samples for this research.

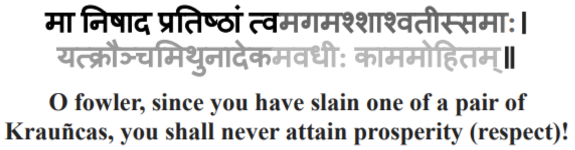
The dataset was initially introduced in the paper titled "Itihasa: A large-scale corpus for Sanskrit to English translation." This paper highlights the contributions of two authors who translated the texts from the two primary sources. The first author, Manmatha Nath Dutt, completed the translations in 1890, while the second author, Bibek Debroy, performed translations in 2010.

3 Dataset info

1. The dataset in sanskrit language was generated using an OCR to extract text from the documents, there were four and nine volumes respectively in which The Ramayana and The Mahabharata translations were published in. The verification of the translations was also carried out manually.In total, it was collected a corpus of 19,371 translation pairs from 642 chapters of The Ramayana and 73,659 translation pairs from 2,110 chapters of The Mahabharata.

2. the dataset is called Itihasa which is a term derived from Sanskrit. In Sanskrit, "Iti" means "thus" or "this way," and "haasa" means "happened" or "occurred.". It contains 93,000 pairs of Sanskrit shlokas and their English translations. The original digitized volumes are available here

3. these are extracted from two literary works that hold great cultural and religious significance in India: Ramayana: The Ramayana - is a story about the life and adventures of Lord Rama. It follows Rama's journey to rescue his wife from the demon king. The epic explores themes of righteousness, duty, loyalty, and the triumph of good over evil. The Ramayana consists of about 24,000 verses; and Mahabharata - is an epic that narrates the Kurukshetra War. The epic dives into complex themes such as duty, righteousness, morality,



मा निषाद प्रतिष्ठां त्वमगमश्शाश्वतीस्समाः।
यत्क्रौञ्चमिथुनादेकमवधीः काममोहितम्॥

O fowler, since you have slain one of a pair of Krauñcas, you shall never attain prosperity (respect)!

Figure 1: Example of a Sanskrit text translated into English.

also the nature of life and death. The Mahabharata is the longest epic in the world, with around 100,000 verses.

4. the language it is also described as "the gods' language" because of its wide use in Indian religious literature from the past

5. The dataset presents asymmetry, where the amount of words needed to express equivalent information is comparatively lower in Sanskrit than in English.

4 Related work

Over time, various models have been trained for this task using the dataset. Models such as B2B-tiny, B2B-mini, B2B-small, B2B-medium, and B2B-base have been utilized for training on the complete dataset, which consists of 74k translated texts for training, 11k translated texts for testing, and 6k for validation. These models were employed to explore different architectures and sizes to assess their impact on translation performance. The highest accuracy of 8.89 percents was achieved using the B2B-Base model. Despite the modest accuracy, it represents the best performance obtained among the trained models for this task. It's worth noting that machine translation from Sanskrit to English is a challenging task due to linguistic complexities, low-resource nature, and structural differences between the languages.

5 Approach

For data preprocessing, I initially loaded the dataset from Hugging Face and developed a function that splits the data into six files: three for English data (train, test, and validation) and three for Sanskrit data (train, test, and validation). Each English file contained one sentence per line, with its Sanskrit equivalent located on the corresponding line in the Sanskrit file. The sentences are delimited by spaces. Only the sentences were retained, while the label types such as "en" (English) and "sn" (Sanskrit) were not preserved from the initial dataset, they were used solely to determine the file in which to add each sentence.To accommodate GPU capacity limitations on my machine, I have specifically selected 20k training samples, 4k test samples, and 1k validation samples for this research. After data preprocessing, I decided to utilize the OpenNMT machine translation model. OpenNMT is a versatile machine translation framework that can be applied to various language pairs, including Sanskrit to English, it is also well-established and widely used machine translation framework that has demonstrated strong performance on various translation tasks. Its effectiveness and accuracy in translating text made it a reliable choice. I initially trained the model for 1000 epochs and observed that the accuracy continued to increase, while the training time remained relatively short. Therefore, I decided to train it for more epochs. By choosing to train for 2000 epochs, the situation remained similar, but with a training time of 10 minutes. Therefore, I increased it to 3000 epochs. After training for 3000 epochs, I noticed that the accuracy reached a plateau around epoch 2500, with a maximum achieved accuracy of 22.4 percents.

6 Limitaions

The difficulty of such a task and its limitations arise from the complexity of the Sanskrit language. Sanskrit presents complex verb forms, noun declensions, and noun genders, making it challenging for models trained on languages with simpler structures. Additionally, Sanskrit presents numerous homonyms and polysemous words, leading to ambiguity in translation. Disambiguating the correct meaning within context can be hard to achive for models. Moreover, the availability of resources for Sanskrit is limited, which further compounds the challenges faced in this task.

7 Conclusion and future work

The primary objective of this paper was to analyze a dataset that consists of translated texts from Sanskrit to English, also to use a pre-trained model. An improvement that could be considered in future iterations is to perform a more extensive preprocessing of the English data by removing punctuation marks. This is because these punctuation marks were still present in the dataset and could potentially impact the translation accuracy. By applying a thorough preprocessing step to eliminate punctuation, the model could benefit from cleaner and more consistent input, potentially leading to improved translation quality. Also, exploring the training of AI models such as MarianMT or VisualTextTransformer could be another solution. These models have shown promising results in machine translation tasks and may offer enhanced capabilities compared to the OpenNMT model.

References

Robert Stephen Paul Beekes. 1995. Comparative IndoEuropean Linguistics. Benjamins Amsterdam.

Itihasa: A large-scale corpus for Sanskrit to English translation.Rahul Aralikatte, Miryam de Lhoneux, Anoop Kunchukuttan, Anders Søgaard.

Improving Neural Machine Translation for Sanskrit-English. Ravneet Punia, Aditya Sharma, Sarthak Pruthi, Minni Jain