**Extremely Low Resource Neural Machine Translation from**

**Sanskrit to English**

**Vijay Rayavarapu, Nikita Salkar, Nissar Ahmed, Shreeyash Yende**

[rayavarapu.v@northeastern.edu](mailto:rayavarapu.v@northeastern.edu), [salkar.n@northeastern.edu](mailto:salkar.n@northeastern.edu),

[pinjari.n@northeastern.edu](mailto:pinjari.n@northeastern.edu), [yende.s@northeastern.edu](mailto:yende.s@northeastern.edu)

Khoury College of Computer Science, Northeastern University  
Boston, MA

**Abstract**

Machine Translation(MT) has been a classical problem in the field of Natural Language Processing. Due to the sequential nature of the process, recurrent neural networks and long short term memory networks have been used in the past. However, RNNs and LSTMs, due to the restrained sequential processing are time consuming. Further these architectures are dependent only on the previously seen states. So, we used the Transformer architecture that accepts the input independent of the prior and processes sentences as a whole rather than sequentially. Transformers also entail the concept of self attention and multi-headed attention wherein similarity scores between words within the same sentence are computed. With a dataset of 38 thousand sentence pairs of english and sanskrit, our model achieved near state-of-the-art result with a BLEU score of 9.119.

**Introduction**

Machine Translation (MT) is the task of automatically

converting one natural language into another, preserving

the meaning of the input text, and producing fluent text in

the output language. It is one of the oldest subfields of Artificial Intelligence research. With the recent shift towards large-scale empirical techniques there have been significant improvements in translation quality. Research over the years has led to broadly 4 types of machine translations.

Rule Based Machine Translation (RBMT), Statistical Machine translation (SMT), Hybrid Machine Translation (HMT) and Neural Machine Translation (NMT).

In Rule Based Machine Translation, the system generates the translations of the input sentences based on morphological syntactic and semantic analysis of the 2 languages. Although this approach does not require bilingual data, it relies on good dictionaries and manually

set rules.

Statistical Machine Translation which does use bilingual text corpora along with phrase-based translations which translate sequences of words or phrases where the lengths may differ. There are several types of statistical-based machine translation models which are: Hierarchical phrase-based translation, Syntax-based translation, Phrase-based translation, Word-based translation. However, it does not consider the context which often results in low quality translations.

Hybrid Machine translation is, as the name suggests, a mix of Statistical and Rule based machine translation. Although its utilization of translation memory makes its output quality more successful, it requires a lot of editing and human translation. There are various approaches to HMT including multi-engine, statistical rule generation, multi-pass and confidence-based translation.

Neural machine translation also requires a bilingual corpus and uses evidently, neural networks to learn a statistical model for machine translation.

In this project we use a neural machine translation model, namely the Transformer.

**Low Resource Setting**

Neural machine translation models feed on a large set of data to learn the nuances of the languages and achieve state-of-the-art results. Machine translation models in the past, especially those that have surpassed human-level performance[2] have relied on availability of large-scale parallel corpora; collections of sentences in both the source language and corresponding translations in the target language of around at least 2 million sentences and corresponding translations.

However, in a realistic setting the vast majority of languages don’t have such resources.

Facebook recently achieved a major milestones wherein

they developed a novel approach that combines

iterative back-translation and self-training with noisy channel decoding, to build an English-Burmese MT

model.[3]

In back translation we are effectively appending whatever output you get which is translations of the source language to your original dataset and iterating through this process again.

Another research conducted at MIT's CSAIL focused on

automatic decipherment of lost languages in an ultimate

low resource setting.[4]

Our project falls in the same category of being extremely low on the resources since we were only able to gather so much.

**Motivation**

We had mainly 3 reasons for choosing to translate from Sanskrit to English.

* After having gathered our data and knowing it was insufficient, we wanted to evaluate the efficiency of transformers in such a low resource environment.
* Also, there are a vast number of texts written in Sanskrit that are yet to be translated for the common public. At the onset of our project we envisioned building a state of the art model that could translate these texts.
* And our third driving factor was the fact that Google

translate itself does not translate to or from sanskrit.

**Data Collection and Preprocessing**

Datasets for Sanskrit to English translation were not readily available. There were no free existing datasets such as the WMT. So, we curated our own dataset.

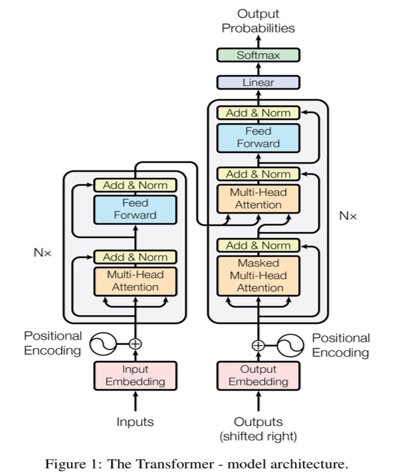
We crawled various websites in search of quality sanskrit literature with available translations in English.

After much perusal, we scraped Holy texts including the Ramayana[5], Vedas, Upanishads and the Bible along with their corresponding English translations. We then generated Sanskrit - English parallel sentence pairs.

The total size of our dataset spanned 38 thousand sentence pairs.

PREPROCESSING data for the transformer: NISSAR

**Architecture**

****

**Experiments**

Our first step was to get an estimate of how well the Vanilla transformer performed.

|  |  |
| --- | --- |
| Number of Encoder/ Decoder layers |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**Results**

**Conclusion**

**References**

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention is all you need", https://arxiv.org/pdf/1706.03762.pdf

[2] CUBBITT

[3]

https://ai.facebook.com/blog/recent-advances-in-low-resource-machine-translation/

[4] https://thenextweb.com/news/new-mit-algorithm-automatically-deciphers-lost-languages

[5] https://www.valmikiramayan.net/

[6]

[7]

[8]

[9]

[10]

[11]