Towards Santali Linguistic Inclusion: Building the First Santali-to-English Translation Model using mT5 transformer and data augmentation.

by

Syed Mohammed Mostaque Billah
20101057
Ateya Ahmed Subarna
23341089
Sudipta Nandi Sarna
20101257
Ahmad Shawkat Wasit
20101398
Anika Fariha Chowdhury
20101042

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
September 17, 2023

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Sudipta Nandi

SUDIPTA NANDI SARNA

20101257

Ateya

ATEYA AHMED SUBARNA

23341089

AnikaFariha

ANIKA FARIHA CHOWDHURY

20101042

Shawkat

AHMAD SHAWKAT WASIT

20101398

Mostaque

SYED MOHAMMED MOSTAQUE BILLAH

20101057

Approval

The thesis/project titled "Towards Santali Linguistic Inclusion: Building the First Santali-to-English Translation Model using mT5 transformer and data augmentation." submitted by

1. Syed Mohammed Mostaque Billah (20101057)
2. Ateya Ahmed Subarna (23341089)
3. Sudipta Nandi Sarna (20101257)
4. Ahmad Shawkat Wasit (20101398)
5. Anika Fariha Chowdhury (20101042)

of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September, 2023.

Examining Committee:

Supervisor:
(Member)

Dr. Farig Yousuf Sadeque
Assistant Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor Dr. Farig Yousuf Sadeque sir for his kind support and advice in our work. He helped us whenever we needed help.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Abstract

Around seven million individuals in India, Bangladesh, Bhutan, and Nepal speak Santali, positioning it as nearly the third most commonly used Austroasiatic language. Despite its prominence among the Austroasiatic language family's Munda subfamily, Santali lacks global recognition. Currently, no translation models exist for the Santali language. This paper aims to remove Santali from the NPL spectrum. We aim to examine the feasibility of building Santali-English translation models based on available Santali corpora. This paper successfully addressed the low-resource problem and, with promising results, examined the possibility of using the Santali language. We think that our study will open the door for further exploration into Santali-English machine translation.

Keywords: Parallel corpus, Machine translation, Neural Machine Translation, Low resource language, Aligner, Transfer learning.

Table of Contents

Chapter 1

Introduction

With over seven million native speakers from India, Bangladesh, Bhutan, and Nepal, the third most widely used Austroasiatic language is Santali. Since Santali had no written language before the nineteenth century, all shared information was transmitted verbally from one generation to the next. Before 1860, European missionaries, folklorists, and anthropologists like A. R. Campbell, Lars Skrefsrud, and Paul Bodding had worked on transcribing Santali using Bengali, Odia, and Roman scripts. As a result of these efforts, we now have Santali dictionaries, adaptations of folktales, and studies on the language's morphology, syntax, and phonetic structure. Although Santali holds the distinction of being the predominant language within the Munda subfamily of Austroasiatic languages, it has not garnered significant global recognition, resulting in its classification as a low-resource language. Like numerous other languages, Santali has received minimal attention in modern NLP research, which primarily concentrates on approximately 20 out of the 7000 languages spoken globally. As a consequence, Santali remains largely unstudied in the field of NLP. Low-resource languages are primarily those that lack extensive monolingual or parallel corpora,in addition to the manually generated language resources required for creating statistical NLP applications. In order to attain state-of-art outcomes in a number of language pairs, neural machine translation (NMT) models have been developed taking the help of recent developments in deep learning. NMT, or Neural Machine Translation relies heavily on vast amounts of data for accurate translations, it is data-intensive technology. These models are built with neural networks, which are capable of finding intricate patterns in data. However, their effectiveness is closely tied to the quantity as well as the quality of the data that is trained. In the case of a lack of data, NMT models may struggle to identify patterns and produce flawed translations. For these models to be trained properly, a substantial amount of sentence pairs need to be provided, which are high in quality. In fact, the absence of such a corpus has a negative impact on the models' performance. In other words, NMT systems may fail to attain desirable results for low resource languages as it is a data hungry technology. Since Santali is a low-resource language spoken by a minor group of people, getting a parallel corpus is quite challenging. That is why our goal is to find appropriate machine translation algorithms and tools for dealing with low resource languages, such as data augmentation and transfer learning, in order to achieve better outcomes. There was a time when a substantial amount of data in the form of parallel corpus was necessary to train a translation model. However, with the advent of newer transformer models, the reliance on data has diminished some-what. Consequently, we wanted to assess whether the available internet resources would suffice to create a translation model for the Santali language, which plays a crucial role in the communication of a community. This thesis allows us to gauge the potential for work on a language with limited resources. If we achieve promising results in this domain, we plan to seek monetary funds from government or private organizations to advance this approach. In short, our objective is to determine if the current conditions are favorable for working with the low-resource Santali language in the field of machine translation. Furthermore, we aspire for our work to serve as the gateway to the realm of Santali machine translation.

1.1 Problem Statement

Despite the enormous progress made in the realm of machine translation, many minority languages are still beyond the reach of this technology. They have not yet been brought under the radar of Machine Translation. Santali is one such language, spoken by seven million people, making it one of such. Therefore, still there hasn't been any machine language tool for this particular language. The revolutionary innovation of Neural Machine Translation in this field has brought massive success in terms of almost human-like translation. However, NMT usually overfits and shows disappointing performance for small datasets. The issue with low resource machine translation is that there is insufficient annotated data for training a machine translation model for a particular language pair. Santali is an endangered, low-resource language spoken by a limited population; as a result, few resources are accessible for its translation. The scarcity of annotated data for Santali-English machine translation can result in poor translation fluency and accuracy, especially for idiomatic expressions, culturally-specific notions, and other language-specific aspects. Along with the data insufficiency, for the language Santali, there is neither any benchmark tokenizer model nor any model that translates santali to any language. The goal of this paper is to test the maturity of the Santali language and to observe its readiness to enter the world of nlp by releasing the first ever santali translation model based on biblical datasets and release Santali tokenizer model.

1.2 Aims and Objective

Primarily our goal in this thesis will be to develop a reliable translation model for Santali-English and English-Santali Language translation.Some other objectives are:

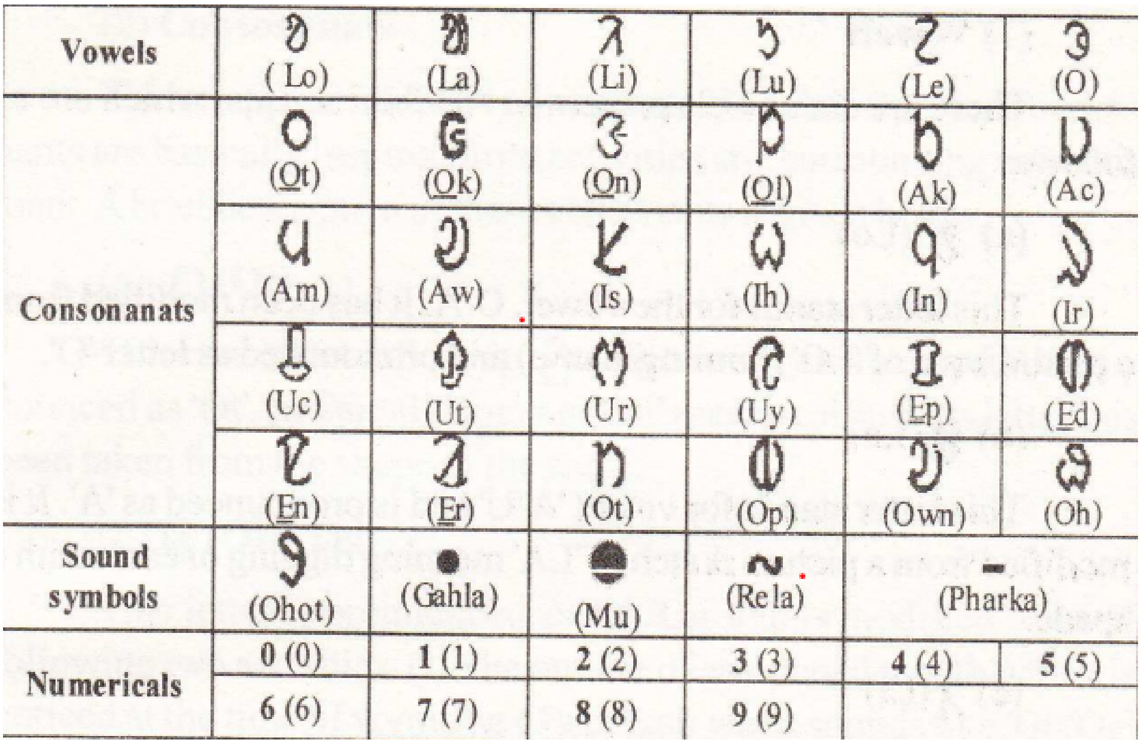
- building a quality dataset for Santali-English translation.
- testing low-resource NMT tools like transfer learning and data augmentation.
- testing multiple models (RNN, NMT, and Transformer) and comparing the performances.

Chapter 2

Background

The Santals are the most numerous of the plain-land ethnic minorities in the northern section of Bangladesh, but they are mostly concentrated in India with a few dispersed populations in Nepal and Bhutan. Apart from Bangla, Bangladesh has roughly 33 languages that are spoken here. These languages are divided into four distinct and significant language families. One of the main languages spoken by Bangladesh's ethnic minorities is Santali, which is a language from the Munda group of the Austro-Asiatic language family with a strong oral culture extending back to ancient times. In areas where Santali is spoken, the literacy rate ranges from 35 to 45%. The Santal people have ancient origins and a deep belief in their oral history and culture. Oral traditions, such as stories, songs, and rituals, have been used for generations to pass along knowledge, customs, and cultural heritage. The Santals had a strong relationship with nature, as evidenced by their mythology and ceremonies that honored natural elements such as woods, rivers, and animals. A large number of words are derived from the natural sounds used by the common or rural Santals in their day to day life. The letters of Ol Chiki scripts are also inspired from the surrounding natural environment as well as the daily used materials which include hills, trees, birds, bees, plough, sickle etc [**18**].

According to Eftakhar(2019)[**16**], the Santali script comprises a total of thirty letters along with five fundamental diacritics, which include an additional set of six primary vowels and three extra vowels. Despite its ancient roots, the Santal language lacked a written form for much of its history. Because they lacked a script, their rich oral traditions were at risk of fading away over time. As the world evolved, the necessity to protect and promote their language and culture became more evident. During British administration, this language was written in the Roman script. The Romanized-based Indian stream refers to the Santal language being written in the Roman alphabet. The Santal language was first documented using this writing system during the British colonial era. Because missionaries and officials could easily use and be familiar with the Roman script at the time, it was a practical decision. Before the 1854 period there was no authorial form of Santali literature. But from 1855 to 1889 in the Missionary Period the first Bible and Christian preaching started in East India, where Santals were also included, especially Santals at Bengal, Bihar and Orissa. As a result, to cross the language barrier missionaries started to learn Santali, practiced the literature and published a book. The first people to recognize that Santali had infused the language with a powerful living energy were the missionaries. This writing system, meanwhile, has certain drawbacks since it might not fully capture the distinctive phonetics and sounds of the Santal language. Santali has also been written in other Indian scripts, including the Bengali script used in West Bengal and the Oriya script used in Odisha. Because of the parallels between Santali and Bengali, many educated Santali writers prefer to write it in Bengali. The reason for this is due to the parallels in the use of phonetics but it was not a popular approach. The Santali language is thought to be older than the Aryan languages. The Santal script is a relatively new development. Santali did not have a written language until the twentieth century, when it employed the Latin or Roman, Devanagari, and Bangla writing systems. Some saw the necessity for a new script that could represent themselves freely, not by adopting foreign or some dominant language scripts, but the script of their own that they would be proud of, as Roman, Bengali, Devanagari, and Oriya were being utilized widely for writing Santali. The distance between the dreams and reality was enormous. However, the emergence of a visionary named Pandit Raghunath Murmu, born in 1905 in the village of Purnapani in the Mayurbhanj district of Odisha (formerly Orissa), was a turning point in the history of Santali. Pandit Murmu while being a teacher, he was also a social reformer who cherished his community from within and its language. Pandit Raghunath Murmu received divine revelation in the early 1920s that encouraged him to construct a script that perfectly represented the phonetics and unique sounds inside the Santali language. In 1925, "OL CHIKI" emerged as a result of Pandit Raghunath Murmu relentlessly putting effort in order to improve the expression of his mother tongue and his passion for excellence which is unique to santal people. Ol Chiki became a developed script that became this language's savior to inspire and maintain a sense of unwavering confidence, unending pride, rebirth , and solidarity among Santals living all over the world. For the entirety of the Santal community this groundbreaking, historical invention helped them to write their words in simpler and easier methods with a great deal of self assurance. With the help of Ol Chiki script provided to the Santal Community they were able to enforce their own methods of written expression and form. It was first created in 1925 and since then it has been changed to 30 characters including 6 vowels and 24 consonants. By keeping the sense of identity and dignity of the people from Santal, The Ol Chiki script attempted to put an effort to reflect the complex structure of phonology and its distinctive sounds of Santali language.The creation of the Ol Chiki script transformed the social and cultural context of the Santal people. The Santals could now preserve and archive their folklore, history, and literary works for future generations because of the written expression. Murmu made significant contributions to the Santhal community via his pen and writings. He wrote approximately 150 books on a variety of topics including grammar, novels, theatre, poetry, and short tales in Santali using Ol Chiki as part of his extensive programme [**21**].



Vowels	ᱚ (Lo)	ᱟ (La)	ᱤ (Li)	ᱩ (Lu)	ᱮ (Le)	ᱳ (O)
Consonanats	ᱛ (Ot)	ᱜ (Ok)	ᱝ (On)	ᱞ (Ol)	ᱠ (Ak)	ᱡ (Ac)
	ᱢ (Am)	ᱣ (Aw)	ᱥ (Is)	ᱦ (Ih)	ᱧ (In)	ᱨ (Ir)
	ᱪ (Uc)	ᱫ (Ut)	ᱬ (Ur)	ᱭ (Uy)	ᱯ (Ep)	ᱰ (Ed)
	ᱱ (En)	ᱲ (Er)	ᱴ (Ot)	ᱵ (Op)	ᱶ (Own)	ᱷ (Oh)
Sound symbols	ᱽ (Ohot)	ᱹ (Gahla)	ᱸ (Mu)	ᱻ (Rela)	ᱼ (Pharka)	
Numericals	0 (0)	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)
	6 (6)	7 (7)	8 (8)	9 (9)		

Figure 2.1: Ol Chiki script from "Endangered Cultures And Languages In India" by Basanta Kumar Mohanta

Santali literature is divided into two categories: primitive literature based on oral tradition and modern literature written by educated Santals. Myths, folklore, and folktales were passed down orally through generations in early writing. These stories offer insight into the Santal belief in gods and spirits, as well as a vivid depiction of their primitive life in forest settlements. "Hor-ko-ren mare Harprarn-ko-reak Katha" and "Kherwal-Vamsa Dharam-Puthi" are two important works in Santali literature. The former comprises the Hor or Santal people's traditions and folklore, gathered and told by a Santal guru named Kolean (Kalyan). Ramdas Majhi Tudu compiled and composed the later, which contains rich information about the customs, religion, and social culture of the Santals and other linked Kol people. In addition to these myths, religious customs, and practices, there exists an extensive collection of Santal folklore. These narratives predominantly revolve around the Santal faith in the Bongas or deities, offering a vivid portrayal of the ancestral existence of the Santal community within their rural jungle settlements[**21**].

While a universal language facilitates communication among individuals from diverse ethnic backgrounds, it carries the perilous prospect of contributing to the gradual erasure of other languages. As each language fades away, a portion of humanity's exceptionally diverse cultural heritage, encompassing countless insights into life, is inexorably relinquished. UNESCO, in its pursuit of identifying the root causes of language endangerment, has identified dwindling speaker populations, prejudicial governmental policies, illiteracy, and the inadequacy of language education as significant factors. Yet, a pivotal concern lies in the influence of electronic media and the internet, both of which play pivotal roles in the endangered status of languages. Their report on this matter paints a disconcerting picture, suggesting that by the year 2100, a mere half of the languages existing today will endure. Social scientists contend that the demise of a language is not merely the extinction of a dialect or linguistic form but also represents the demise of a crucial facet of culture, marking a profound loss for humanity itself. Literature is often regarded as a reflection of society, and it can effectively fulfill this role when it is crafted using the language unique to a particular community or group and written in their specific script. Expressing the full meaning of all words in a given language or pronouncing them accurately becomes quite challenging when attempting to do so in a foreign language unfamiliar to that community or group. Consequently, for the enhancement of a specific language and its literary tradition, it is essential for it to possess its own script [**10**]. Language and its representation hold paramount significance because every small community possesses its unique traditions and culture, with language being one of its most vital cultural components. Due to the pervasive influence of dominant languages and cultures, minority groups and their languages face significant challenges in preserving their existence. In the contemporary context, Santal literature stands as a vibrant phenomenon, manifesting itself through various literary forms such as drama, poetry, novels, historical narratives, folklore, riddles, and even newspapers. Santal writers assume charismatic roles, adept at not only framing traditional cultures as sources of inspiration but also addressing the various issues that Santals encounter as they navigate a globalized world [**16**]. When we examine Santal Literature in the present context, we observe that it continues to advance steadily. While its forms have evolved, its inherent value has remained constant. From this standpoint, it is unequivocal that Santali Language and its literature have developed considerably. If we can contribute economically to this ongoing advancement, its progress will extend even further[**23**].

Chapter 3

Literature Review

3.1 Sentence Alignment system

In this paper [**6**], several aligning methods have been briefly reviewed and evaluated running on the same SMT systems trained on the outputs from the alignment methods.The experiment had shown that using a better sentence alignment method can result in gaining 0.5 to 1 BLEU points.
Parallel corpora is an essential element in Natural Language process applications.So, growth of the data-dependent NLP methodologies is hindered by the absence of such parallel corpora for specific languages addressed as minority language.To reduce the ambiguity word correspondences of each pair the segments can be handled by a technique called "Sentence Alignment" which indicates correlation between two segment.
In this paper, the authors briefly explain different tools for sentence alignment using two parallel corpora. Then they compared the performances for those tools addressing the fact that theory were trained upon same dataset using the same SMT model.They have evaluated five sentence alignment tools.

The Gale- Church Approach [1]works on the principle that two sentences are highly likely to be of almost same length.So,the algorithm looks for sentences that are comparable in length..
Bilingual Sentence Aligner[2] extended the above mentioned concept integrating the idea for word mapping. Firstly, based on sentence length the corpus is aligned then this is passed as the training set to a lexical model.These alignments comprise 1-to1 mapping with high precision.
Though **Hunalign**[3] is almost identical to Moore's algorithm, with a difference of dictionary based substitution. This results in speed optimization. However, the memory consumption is not well handled.
Gargantua[4] replacing the second step of Moore's algorithm this alignment method uses two steps for the second pass. In the first step,they collect 1-to-1 alignments dynamically and then in the second step they combine these with unaligned sentences. Thus they obtain 1-to-many and many-to-1 alignments.
Bleualign[5] needs the source text to be automatically translated. This algorithm uses two passes, firstly to find 1-to-1 alignments that maximizes the total score and then further 1-to-1, many-to-1 and 1-to-many alignments are added in the second pass.

In the paper, on the same dataset the same SMT system had been trained just with different sentence alignment tools. The result of the experiment shows, Bleualignand Gargantua aligned SMT models performed better than others on the Urdu-English dataset . And Bleualign and Hunalign performed better in the French English data. The research outcome of this paper exhibits the importance of choosing a more advanced sentence alignment algorithm.

3.2 Neural Machine Translation

In spite of being a very powerful model, deep neural networks have some limitations that they cannot overcome. The reason why DNN models face such a problem is because the DNN network works with vectors of predetermined dimensions. That means the model needs to know the number of input and target vectors, which is a problem in the case of translation. “Sequence to Sequence Learning with Neural Networks” [8] by Ilya Sutskever proposes a model where it takes the input ”XYZ” and can predict “ABCD.” It will start predicting and keep predicting until it predicts the end of the sentence. Another claim that this paper makes is that using reversed LSTM on source sentences gives better output due to the introduction of new dependencies. This model uses RNN, a generalization of a feedforward network. By iterating the certain equations, the model predicts outputs from inputs. The conditional probability of p(y1,..., yT —x1,..., xT) is calculated by LSTM, where “x” is input and “y” is output, and T and T’ are the sizes of input and output, which can differ from each other. The model employs two LSTMs and trans both on the language pair at the same time. It also uses multilayer (four-layer) LSTM for performance enhancement, and it predicts target language from a reversed input language to increase dependencies. They have used the WMT ’14 English to French dataset to train their model. This paper saw a jump of 4.7 BLEU points just by reversing the sentence. Another finding of the paper is that, though LSTM has a short memory, it works very well for long sentences. his model outperforms the previously state of the art model by 0.5 BLEU.

The old NMT models, such as phrase-based systems, uses a vector that has fixed length to represent a source sentence, which may lead to loss of information for sentences that are long. Overcoming this limitation, the proposed model [7] encodes the source sentence into a sequence of vectors, allowing for a dynamic representation of the sentence then the previous approach. The model also uses a bidirectional RNN encoder-decoder that adaptively selects subsets of vectors during translation, rather than relying on vectors that is fixed in length. The decoder uses a probability based function that takes into account the previous translations, the hidden states, and a context vector to guide the translation process. The bidirectional RNN allows for an annotation of each word that includes not only the preceding words, but also the following words in the sentence by merging the forward and backward hidden state. RNN usually focuses on outputs that are calculated in recent stages.. As a result, the annotation of a word will focus on the surrounding words of the particular word. The model uses English-to-French parallel corpora to train its model and finds notable improvement over traditional phrase-based translations. The paper demonstrates that, while longer sentences show better improvement, this model provides a better translation result than any other traditional encoder-decoder system, regardless of sentence length.

The English-to-German translation model using attention-based NMT produced a new state-of-the-art model at that time with an improvement of 1 BLEU score over the past state-of-the-art model. This paper has used two approaches: an approach that attends to all source words (the "global approach") and another method that attends to a subset of source words at a time (the "local approach"). Global attention Approach: The paper **[9]** tried to simplify its architecture from the STOA model that it was built on. They have used stacking LSTM architecture. At the top LSTM layers, they have used hidden layers in the encoder as well as the decoder. They have used multiple alignments to test for better results. This approach requires the model to go through each source word while translating a target word. That means translating longer documents, paragraphs, or even long sentences can be computationally extensive and impractical at times. That's when they introduce another approach called the "local attention approach." Local Attention Approach: This approach is easier to implement than the global approach. In spite of focusing on every source word, it focuses on a small batch of words. The model creates an aligned position for every target word at a given time. The set of source hidden states of the focused word is then weighted and averaged to make the context vector. To keep things simple, the input-feeding approach is being chosen. This simply means that the model passes its attentional vector as an input to the next step in order to pass the alignment decision history information. Doing so, this paper hopes that the model knows about past alignment decisions. Using 4.5 million pairs of sentences, this paper has observed an improved score of 5.0 and 1.0 on no-attentional NMT models and attentional NMT models, respectively. Their findings are enough to conclude that attentional NMT models are superior to non-attentional NMT models in various cases (translating long sentences).

In contrast to the newer language models, BERT is designed for pre-training deep bidirectional representations from unlabeled text by altering the left and right contexts of all layers together. The two existing techniques for applying pre-trained linguistic representations to downstream tasks are feature-based optimization, which uses a task-specific architecture with pre-trained representations as additional features, and Fine Tuning, which is trained for downstream tasks by introducing minimum task-specific parameters.The main drawback is that current language models are unidirectional that is limiting the architectures which can be utilized in the time of pre-training. These limits are suboptimal for tasks at sentence level and may be problematic when applied to token-level tasks like question answering, where including context from both directions is crucial. BERT is introduced in this paper **[13]** to generate deep bidirectional representations using masked language models and to improve fine-tuning-based approaches. Pre-training involves using two unsupervised tasks to train the model on unlabeled data. Masked Language Modeling masks words in a corpus at random and asks the model to predict the words in a sentence. Bidirectionality is achieved by masking 15 percent tokens in a phrase with a special token and using cross-entropy loss to predict those masked tokens. A binarized Next Sentence Prediction task is generated to comprehend the relationship between two sentences. By training the model with these parameters, it is fine-tuned for the desired downstream task. BERT uses the self-attention mechanism of the Transformer model architecture, which encodes the concatenated sentence pair with self-attention and includes bidirectional cross attention between the two sentences. To demonstrate the significance of BERT's deep bidirectionality, the authors compare two models trained using the same pre-training data, fine-tuning scheme, and hyper parameters such as No NSP, a bidirectional model trained using the masked language model but lacking next sentence prediction, and LTR No NSP, a model trained without the MLM but with a left-to-right linguistic model and lacking NSP. When bidirectionality of the language model is eliminated, it performs worse than the MLM model on all tasks. Though BERT promoted the idea of training large transformer models on large datasets and essentially transformed NLP research, the architecture has several limitations as it is bidirectional and there is no apparent way to train BERT for auto-regressive tasks.

NMT models depend on large databases. Good translation models need huge databases. Multiple studies show that even after using techniques like "zero shot" and "few shot," small corpora cannot produce performance compared to big corpora. Bengali, in spite of being the world's seventh most spoken language, lacked good-quality parallel corpora for English translation. The paper "Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation" [**17**] by Tahmid Hasan presents a new model for Bengali-English machine translation that significantly outperforms previous models with a BLEU score of more than 9.0. The improvement was achieved by introducing new high-quality data, fine-tuning the sentence alignment process and pre-processing the training data. The paper describes how these methods helped to improve the performance of the Bengali-English machine translation model. The paper discovered that traditional sentence segmentation libraries, such as Polyglot, did not work well on Bengali. Instead, using a rule-based segmentizer like SegTok produced better results. This paper used the extended code of SegTok to add rules for Bengali. Moreover, the paper concatenated the results of three aligners—Hunalign, Gargantua, and Bleualign—to align sentences. Lastly, the paper used another technique called "batch filtering," where they pre-processed data to train a better model. Two types of training data were used: "sentence-aligned corpora" and "document-aligned corpora." The paper has developed its own evaluation method because traditional methods did not work well enough for Bengali. The combination of all these methods resulted in a model that produced more than a 9.0 BLEU score advantage over the previous Bengali to English translation and more than a 5.0 BLEU advantage over the previous English to Bengali translation.

3.3 Low Resource NMT

The Encoder Decoder Neural Machine Translation had been showing promising results,but only on large data. When it comes to the low resource language, NMT fails to show any effective performance. This paper [**11**] proposes a Transfer model for low resource NMT, that significantly improves performance. Their proposed method is to primarily train the model with a high resource language pair and then pass the obtained parameter to the low resourced pair. They are calling the first one as **parent model** and second one **child model**.With this idea, they were able to get an average of 5.6 BLEU on four language pairs with few resources.They also introduced Ensembling and unknown word replacement to enhance the performance.Additionally, they have shown a significant improvement of the SBMT system using the transfer learning model for rescoring. Transfer learning is basically transferring the knowledge from a learned task to any related task. Here basically it passes the learned parameters. As a result the child task needs less training data. This paper incorporates the idea of transfer learning in NMT. Firstly training a NMT model as the parent with sufficient data ,then training another NMT model (*child model)* initializing with the parameters of the parent model. The intuition behind this is, since very limited data is used for the *child* model, it will need justified prior distribution. Finally they modify few of the parameters of the *parent* model,while leaving the rest to be finetuned by the *child* model.The paper also focused on the choice of parent language. They showcase a comparison result between French and German as parent language for Spanish language, where French performs better than German due to the similarity of French and Spanish. Extending this knowledge, they further carried experiments on relatedness of parent class. The results indicates, performance of such models increases for “closely related” parent classes. Transfer learning NMT models performs very well for low resource languages and almost reaches SBMT systems, additionally ,notably improves state-of-the-art SBMT systems on LRLwhen used for re-scoring.

Data augmentation is a popular method used in computer vision-related AI training models with insufficient data. For example, rotating, clipping, or flipping an image would not change its label. So in computer vision, by changing image parameters, multiple instances can be created of the same labeled image. While it improves model accuracy in computer vision, the process of data augmentation can be applied to train low-resource NMT translation models. Synthetical parallel corpora can be created for rare words, and Marzieh Fadee explained an easy way of doing so in her paper, “Data Augmentation for Low-Resource Neural Machine Translation.” [**12**] Sequence-to-sequence translation architectures using LSTM need multiple instances of the same word in different contexts. Creating data by manual annotation is not possible all the time. Creating synthetic data can be one of the most efficient ways to go. This paper proposes identifying words that have occurred less than a threshold of times in a corpus as rare words. Then new sentences are made by replacing semantically similar common words with rare words. If the new sentence is plausible both semantically and syntax-wise, then the new sentence is accepted. The same changes are then made to the translated sentences. Newly created sentences are kept based on probabilistic forward and backward LMs. This model produced 3.2 more BLEU points than back translation.

Gu et al in this paper [**14**] proposed a unique technique called “Universal Neural Machine Translation” (UNMT) to overcome the challenge of Low Resource data for NMT.In their approach, they train a NMTon a large, and then using a smaller dataset they finetuned the model for a specific target language. The intuition behind this approach is that the model can primarily absorb the general patterns and features of language translation that apply to a wide range of languages,then use this knowledge to more accurately translate the particular target language. The authors used Universal Lexical Representation (ULR) and Mixture of Language Experts (MoLE) to enable both word-level and sentence-level sharing, respectively. To evaluate UNMT, experiments on a number of low resource language pairs had been conducted . These included pairs where one language had a significant amount of annotated data and the other had a little amount (for example, English and Estonian), as well as pairs where both languages had a modest amount of annotated data (e.g. English-Lao).The results had shown that UNMT was able to transfer knowledge from high resource languages to low resource languages, further improving performance, and consistently outperformed strong baselines and other approaches for low resource language translation. The authors also carried out ablation experiments to comprehend how various UNMT model components influenced translation quality.They identified that the using both source and target language data during training was important for the model to perform well, and further it is essential to fine tune the model on the target language dataset, since it help the mode for adapting to the specific characteristics of the target language. In addition to its strong performance on LRL pairs, UNMT is able to perform well on HRL pairs, making it a potentially valuable technique for machine translation in general. Overall, this paper presented an highly effective approach for LRL using universal neural machine translation. UNMT was able to perform well in translation even in the absence of a significant amount of labeled data for a specific target language by leveraging the general patterns and features of language translation learned from a diverse dataset of languages.

Extending the concept of model-agnostic meta-learning algorithm, Gu et al. in this paper[**15**],proposed a approach based on meta-learning for low-resource NMT (Meta-NMT). Meta-learning is a type of machine learning method entailing how to learn, or learning the learning process itself. In the context of NMT, this refers to learning how to translate a specific language pair using insufficient annotated data.The underlying idea is to have initialization parameters to be learned previously from a set of source tasks having adequate data,so that with very few training examples the targeted task can be learned. Here in machine translation the intuition is leveraging various high-resource language pairs to learn general patterns and features that are applicable across many language pairs a new translation model is trained on a low-resource languishing the knowledge from those auxiliary languages.They used the universal lexical representation [**14**] to have similar input-output across different languages.The first subtask replicates learning a language specifically, while the second subtask reviews the results.The updated parameters are then evaluated on the second chosen task. To update the meta-model, they used the computed gradient from the previous evaluation that they referred to as a “meta-gradient” .Unlike traditional learning, the model generated through meta-learning is not necessarily a viable model independently. It is, nevertheless, a smart start when training a good model with just a few learning steps.Moreover, this proposed model varies from transfer learning in a sense that,The NMT system is trained using Transfer Learning on a specific source language pair, and then each target language pair is individually fine-tuned using the same system. Overall, this paper presents an effective approach for low resource NMT using meta-learning.Although Meta-learning has been used in the past for many different tasks, but its use in NMT is relatively novel.And the results from this research justifies for this technique to be an effective approach for low resource NMT.

Chapter 4

Dataset Collection, Preprocessing and Analysis

4.1 Data Collection

The data is taken from a website called ’bible.com, and it is currently the biggest dataset that is available all over the internet as the Bangladeshi-Santali data set. The Santali Bible had 66 books and multiple chapters in each book. Among them, ”SEREKO 1” had the highest number of chapters (150). There were several books with only one chapter. We built a customized data crawler to crawl through all the books and all the chapters to fetch all the Santali-English-Bengali parallax corpora that were available for us.

From there, we found 29,651 sections, which may or may not contain more than one sentence. The total number of sentences was 69,086; the total number of words was 6,63,684. The length of the word set was 50,822. The top 10 most frequently occurring words are as follows:

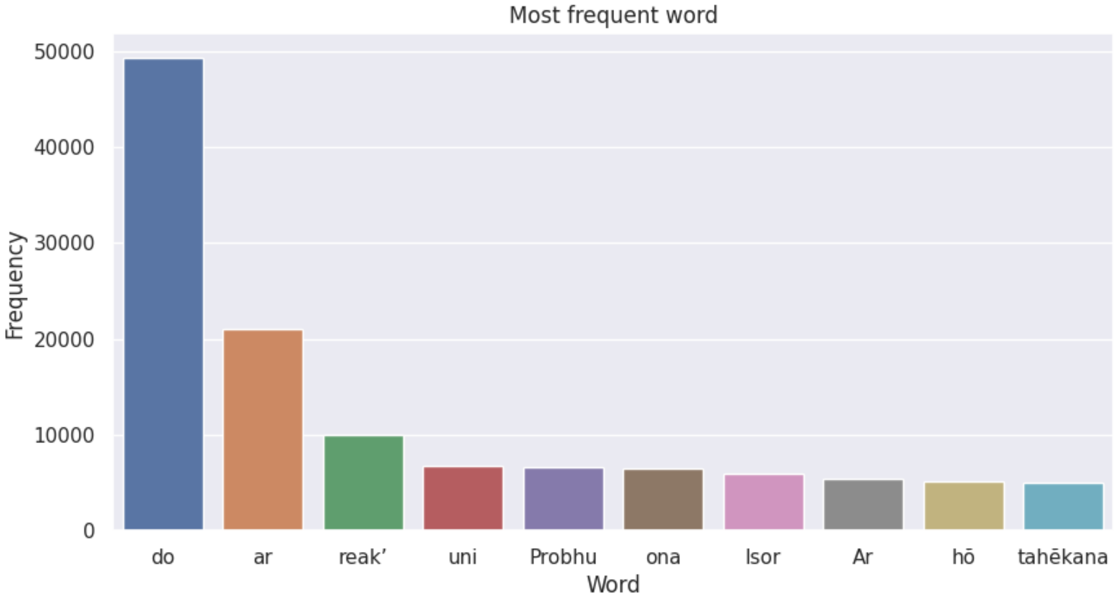


Figure 4.1: Barplot of most frequent words (top 10)

since the Santali Bible database does not contain enough information. Most of the words from the unique word list did not occur more than once. For example, more than half of the words found were present only once. Which is definitely not enough for any model to learn its context. The number of words that occurred only once is 28,525; the number that occurred only twice is 7,142. Here is the list of the top 10 frequencies of the least frequently occurring words in the whole corpus:

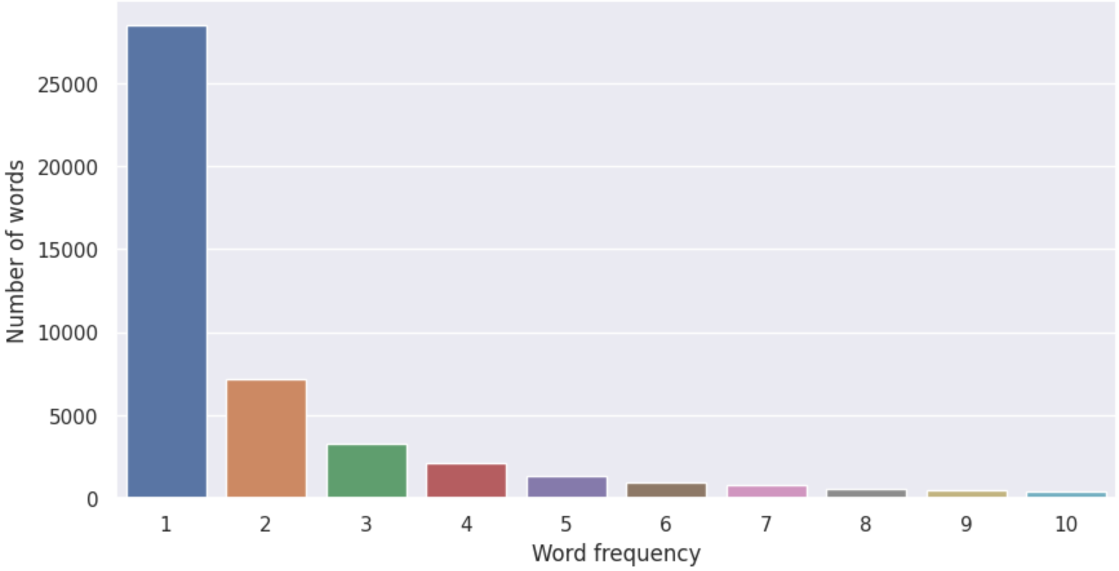


Figure 4.2: Barplot of frequency of least occurring words. (Top 10)

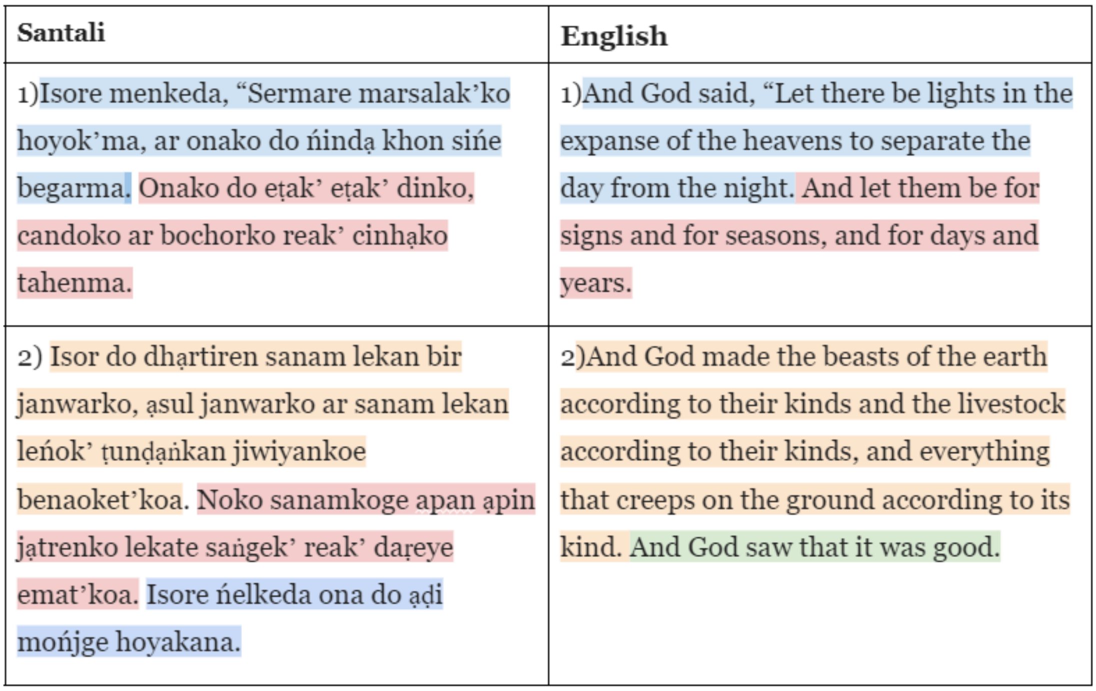
4.2 Data Pre-processing

4.2.1 Data Cleaning

After data extraction, we executed some basic pre-processing techniques, such as the removal of punctuation marks. Then we had to encode the Santali data since it uses special Roman characters. Besides, we had to do some sentence level assignments which had been discussed in data splitting.The data was very clean to begin with, so we did not need to go through rigorous pre-processing.

4.2.2 Data Splitting

Our dataset included sets of sentences as one unit pair.Where two or more sentences were grouped as one sentence having a group of corresponding target sentences, which may or may not be or same length. For example,



Santali	English
1)Isore menkeda, “Sermare marsalak’ko hoyok’ma, ar onako do ńindạ khon sińe begarma. Onako do eṭak’ eṭak’ dinko, candoko ar bochorko reak’ cinhạko tahenma.	1)And God said, “Let there be lights in the expanse of the heavens to separate the day from the night. And let them be for signs and for seasons, and for days and years.
2) Isor do dhạrtiren sanam lekan bir janwarko, ạsul janwarko ar sanam lekan leńok’ ṭunḍạṅkan jiwiyankoe benaoket’koa. Noko sanamkoge apan ạpin jạtrenko lekate saṅgek’ reak’ daṛeye emat’koa. Isore ńelkeda ona do ạḍi mońjge hoyakana.	2)And God made the beasts of the earth according to their kinds and the livestock according to their kinds, and everything that creeps on the ground according to its kind. And God saw that it was good.

We see in example-1 the source unit and the target unit has the same number of sentences, while in example-2, the sentence number varies in the source and target unit. Such lengthy units hardly can perform well in NMT models. So, we broke down the dataset in two parts,namely, ‘same length units’ , ‘variable length units’. For the same length units, we broke-down the units into single sentences and mapped each source sentences to the target sentence. And , on the variable length units, we kept the dataset unchanged. Lastly to ensure, our training corpus has both one-to-one mapped sentences and variable length units, we splitted the same length units and variable length units in 0.8,0.1.0.1 for train, test and validation set and merged them in the final training , testing and validation data.

Algorithm followed in creating the dataset-

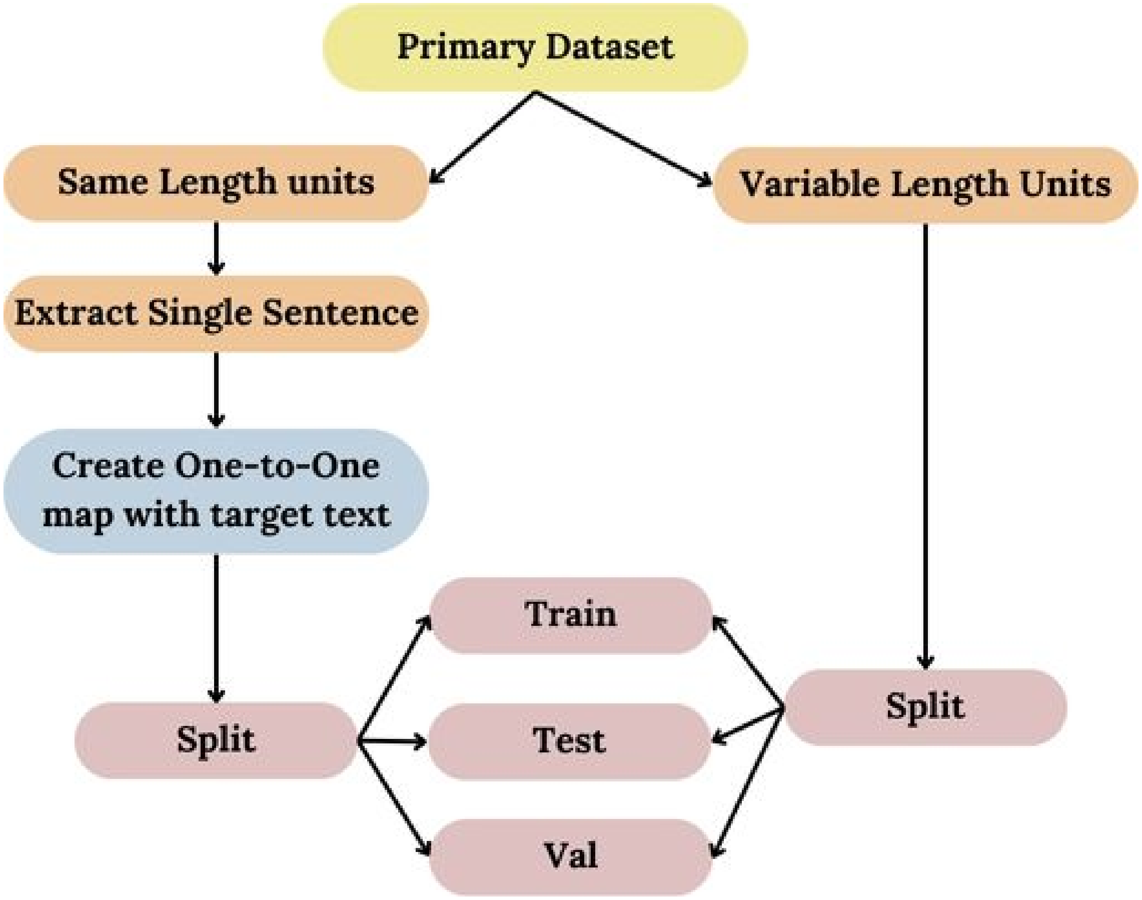


Figure 4.3: Flowchart to create train-test-validation dataset

Doing so, we get-

Dataset	**Primary Text Units**	**One-to-One Sentences**	**Variable Length Units**	**Final Text Units**
Santali-English	29,651	25,153	8,250	33,403
Santali-Bangla	29,166	21,240	12,584	33,824

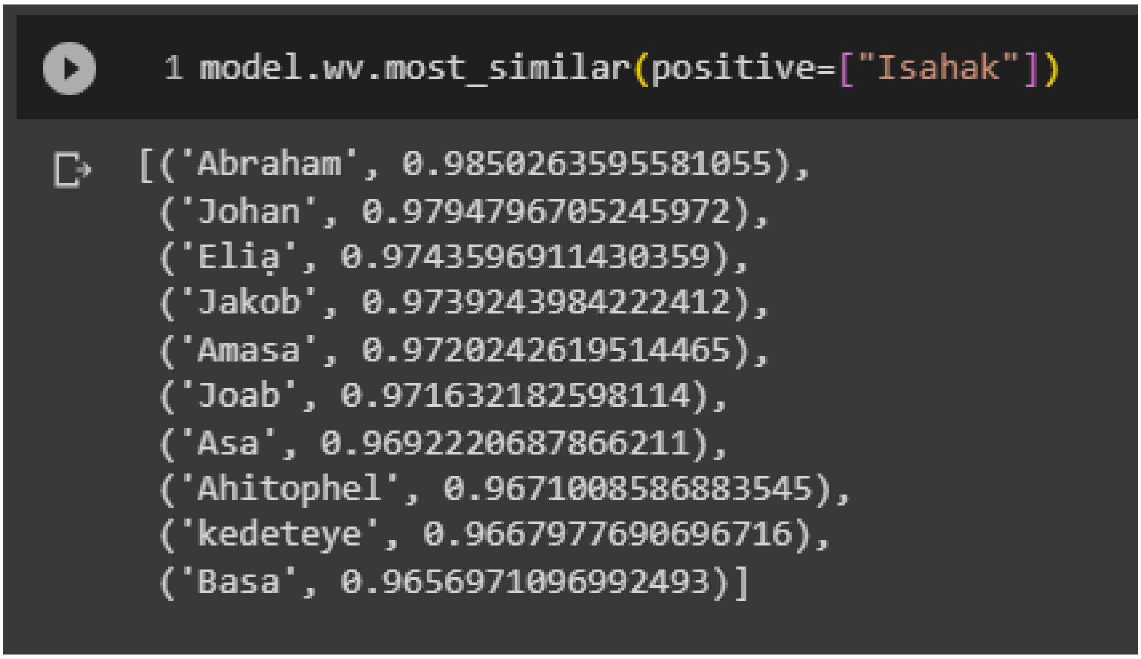
Train-Test-Validation split

Santali-English	One-to-One Sentences	Variable Length Units	Total
Train	20,123	6,600	26,723
Test	2,515	825	3,340
Validation	2,515	825	3,340

Santali-Bangla	One-to-One Sentences	Variable Length Units	Total
Train	16,992	10,068	27,060
Test	2,124	1,258	3,382
Validation	2,124	1,258	3,382

4.3 Word2Vec

Then comes the vectorization of the words. Meaning that words needed to be presented through numerical value in a way that they held some contextual information. The creation of this word embedding was also crucial for us, as from this we could understand whether our data is able to preserve any contextual quality or not. The results were optimistic, as when we wanted to find the most similar words to "Ishak," the prophet, it gave us ten more names who were likely prophets themselves:



```
model.wv.most_similar(positive=["Isahak"])
```

```
[('Abraham', 0.9850263595581055),
 ('Johan', 0.9794796705245972),
 ('Eliạ', 0.9743596911430359),
 ('Jakob', 0.9739243984222412),
 ('Amasa', 0.9720242619514465),
 ('Joab', 0.971632182598114),
 ('Asa', 0.9692220687866211),
 ('Ahitophel', 0.9671008586883545),
 ('kedeteye', 0.9667977690696716),
 ('Basa', 0.9656971096992493)]
```

Figure 4.4: Most simmilar words to "Ishahak" based on cosine similarity. (Top 10)

Our model can also produce a list of most similar words, and it can solve equations to give us words. Such as king + woman + man = queen. Such as:

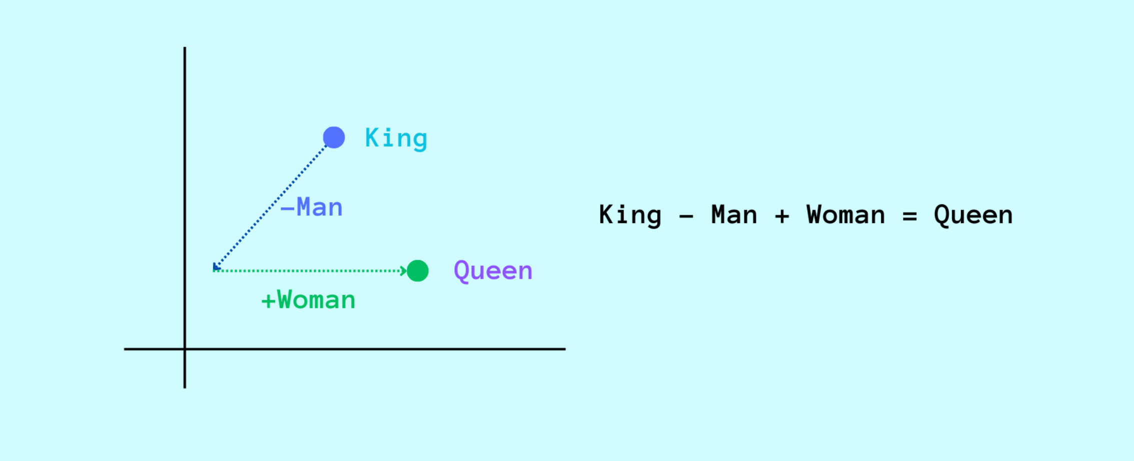


Figure 4.5: Visualizing properties of word2vec with imaginary value

But unfortunately, since we do not possess much knowledge of the language, we cannot show this task in Santali in a way that will be understandable to us.
Our model can also produce a lower-dimensional representation of words with their cosine similarity. Such a graph is shared below that contains ten most similar and dissimilar words. Graph for ”Isor” is shared bellow:

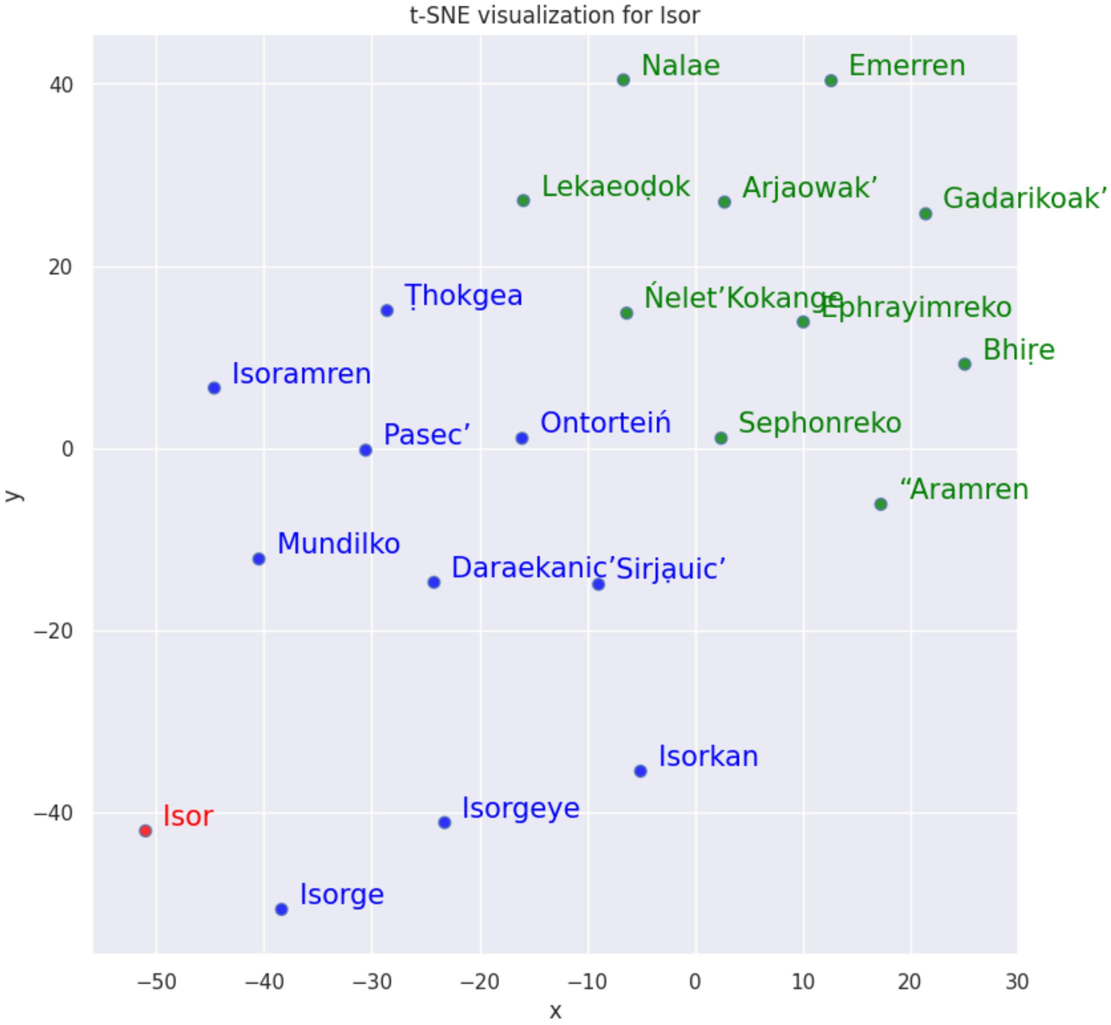


Figure 4.6: t-SNE plot to visualize word vectors in a 2D plain. Red- the source word, Blue - similar words, green- dissimilar words.

Chapter 5

Methodology

5.1 Aligner

The idea of alignment in machine translation denotes the problem of the underlying ambiguity of two words in parallel corpora. It mainly occurs for two reasons:

- Position of the words in a sentence
- Number of unequal words in a sentence.

We can understand the problem better from the example given below:

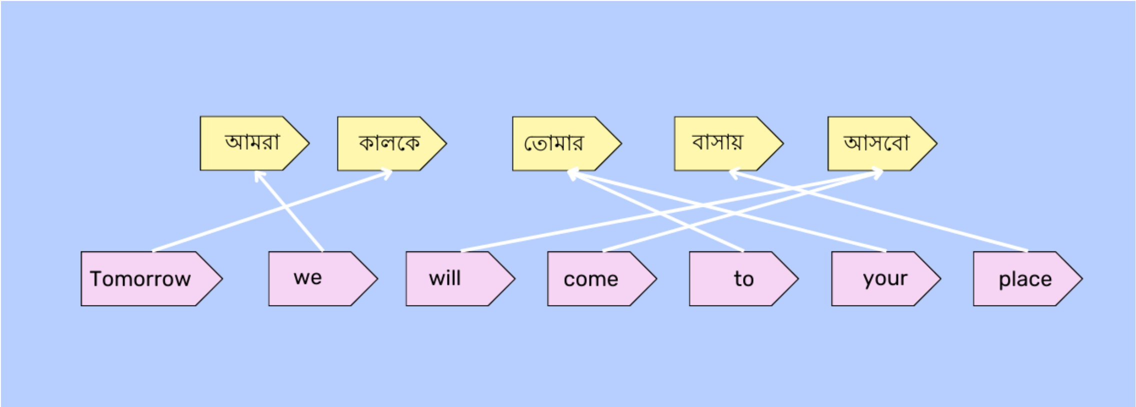


Figure 5.1: Correspondence of word of source sentence and target sentence.

As we see, machine translation more often produces one-to-many mapping problems. In our given example, the first word of the source sentence does not correspond with the first word of the target sentence. Moreover, the number of words in the two sentences is unequal. Though most of the word-translation is one-to-one mapping, words like “will, come” and “to, your” exhibit many-to-one relationships. Over the years, many researchers have delved into this problem in their own unique ways. But the majority of the attempts had one thing in common: “probability”. Most of the papers, in one way or another, have tried to establish correspondence with words based on their probability. The idea quite simply goes like this: every time a word is in a parallel sentence with another word, their probability of correspondence increases. Then, from the source sentence, a word calculates its probability with every other word in the target sentence and gets associated with the word with the highest probability. For example:

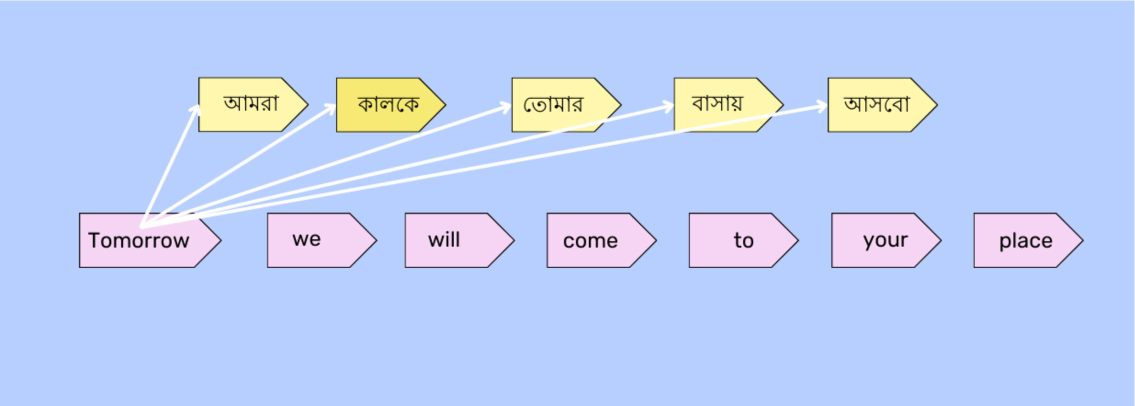


Figure 5.2: Probabilistic Alignment process.

Here, after calculation, we know the corresponding word of tomorrow is the second word from the source sentence.

There are two types of aligners: word-based aligners and sentence-based aligners. Sentence-based aligners are usually used in languages that are not separated by a space after each word. Such as the Japanese language. Example:

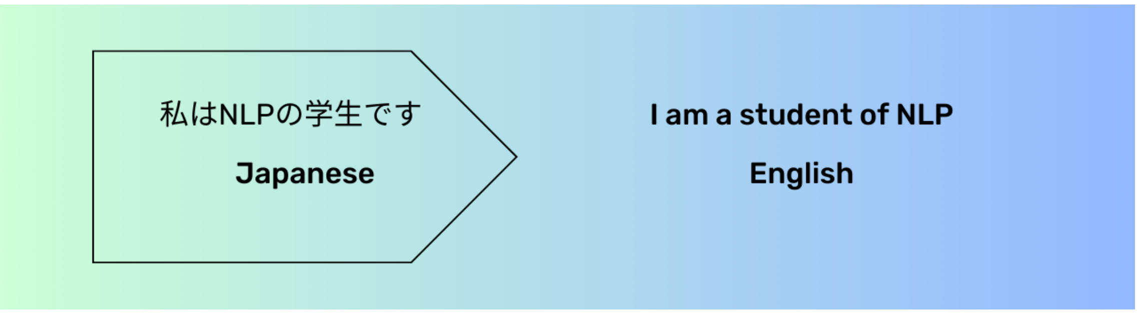


Figure 5.3: Need of sentence based aligners.

In this case, sentence-based aligners can be used. There is another case where sentence aligners come in handy, and that is when parallel corpora do not have parallel sentence pairs. But both cases do not apply to our project, as the dataset we are working with has parallel data. So we worked with word-based aligners.

In our seq-to-seq encoder-decoder model, we have used attention mechanisms to tackle alignment problems. Attention mechanism essentially means a method for words from the source sentence to understand what words to attend to in the target sentence. With the context vector that's been passed from the encoder to the decoder, an attention weight is sent in the latest seq-to-seq models[**24**].

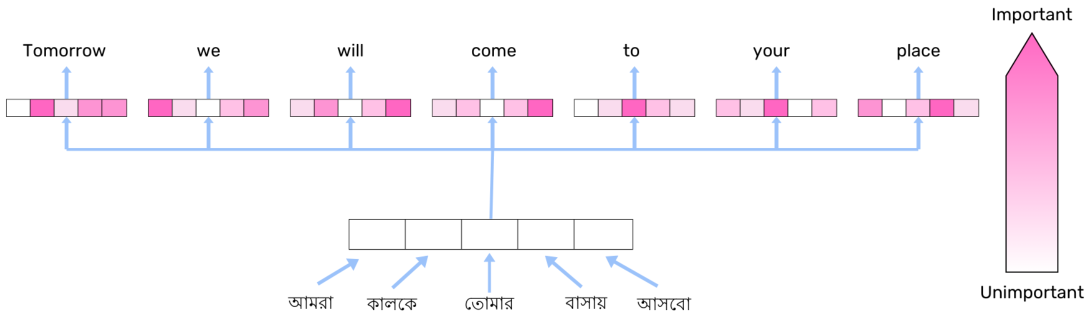


Figure 5.4: Aligner with attention mechanism.

From this architecture, we see that our input and output vectors are fixed in size, but the number of vectors does not need to match, truly preserving the many-to-many nature of machine translation.

Besides using alignment, during data preprocessing we applied a length-based aligner. In our parallel data, we had data points where there were multiple sentences parallel to multiple sentences (sequence not maintained). In such a situation, we tried to look for parallel sentences of equal length, and manually, we saw that our analogy quite worked with the data set. Hence, we were able to further align some sentences with brute force.

5.2 Transfer Learning & Data Augmentation

Neural Machine Translation(NMT) requires a large amount of data for training the model in order to get adequate results. However, in cases of Low Resource Languages (LRL) it is difficult to accumulate such large dataset. Therefore, techniques like Transfer learning, Data Augmentation, and Multilingual Models are used to improve the performances of NMT models. In our proposed model, we have used two techniques combined to improve the performance of our model.

5.2.1 Transfer Learning

Transfer learning involves using a pre-trained model on a related task and fine-tuning it for the specific low-resource NMT task. For instance, models like BERT or GPT, originally designed for tasks like language modeling or text classification, can be fine-tuned for translation. Transfer learning can help leverage knowledge learned from high-resource languages to improve the performance of low-resource NMT. In this paper[**11**], it has been shown how transfer learning improves the NMT model's performance. Moreover, when a parent model is chosen, which is closer to the child model, the model performs better than the non-similar parent model.

For our model, to leverage transfer learning we used MT5 -Small [**22**] as the parent model. MT5 model is an extension of T5 **[20]** which is based on Text-to-Text transformer. mT5 is a multilingual Transformer model that is trained on a dataset (mC4) having text from 101 different languages. MT5 model is designed to support various NLP task such as classification, NER, question answering etc by reaccessing the required task as a sequence-to-sequence task. It has 300 million parameters.

5.2.2 Data Augmentation

Data Augmentation is another technique to enhance the performance of low-resource NMT models. With Data Augmentation, the amount of data is increased. There are three well-known strategies for augmentation.

1. **Back Translation:**Here the text is translated to another language and then translated back into the original language.
2. **Easy Data Augmentation:** This technique includes four methods: synonym replacement, deletion, inserting randomly, and text swapping.
3. **NLP Albumentation :**When there are duplicate values, sentences are shuffled or excluded.

For our model, we used both Back Translation and Easy Data Augmentation technique for both Santali- Bengali and Santali-English dataset.

Santali-English:

For our Santali to English dataset, we augmented the English part of our training corpus. Our primary training dataset contained 26,724 pairs of sets of sentences, where 20k pairs were one-to-one mapped. We labeled this as a 'good dataset' and augmented these 20k sentences. We excluded the sentences, which were not one-to-one mapped, from the augmentation process, since the sentences were too long, and making tiny changes to such long sentences would not affect much to create comparatively newer sentences. From the TextAttack **[19]** framework we used the EasyDataAugmenter class as our augmentation recipe.This class implements-

- **WordNet synonym replacement:** Here a word is selected randomly and then replaced with a synonym
- **Word deletion:** Removing a word randomly.
- **Word order swaps:** Randomly choose two words and swap their position
- **Random synonym insertion:** Select a word randomly and insert it's synonym in a random position

All this is done in one Augmentation method. After augmentation, we had a training corpus of 46,726 sets of pairs of sentences.

Santali-Bangla:

In order to augment the Bangla texts we used bnaug [1] library. Similar to the English text, we extracted the Bangla texts from the pairs and passed it to the bnaug library. We used token replacement and back translation both on each bangla sentences to generate a single augmented instance. For token replacement we utilized random mask-based generation using pre-trained Bengali GloVe10 and Word2Vec11 embeddings. Again, using the libraries backtransaltion method we paraphrased the mask based augmented sentences to paraphrase. Initially, in out training corpus we had 27,060 sets of pairs of sentences, however with augmentation, we generated more 15,000 pairs of sentences. Finally out training corpus had 42,060 sets of pairs of sentences.

The algorithm we followed, creating the augmented dataset for both Bangla and English:

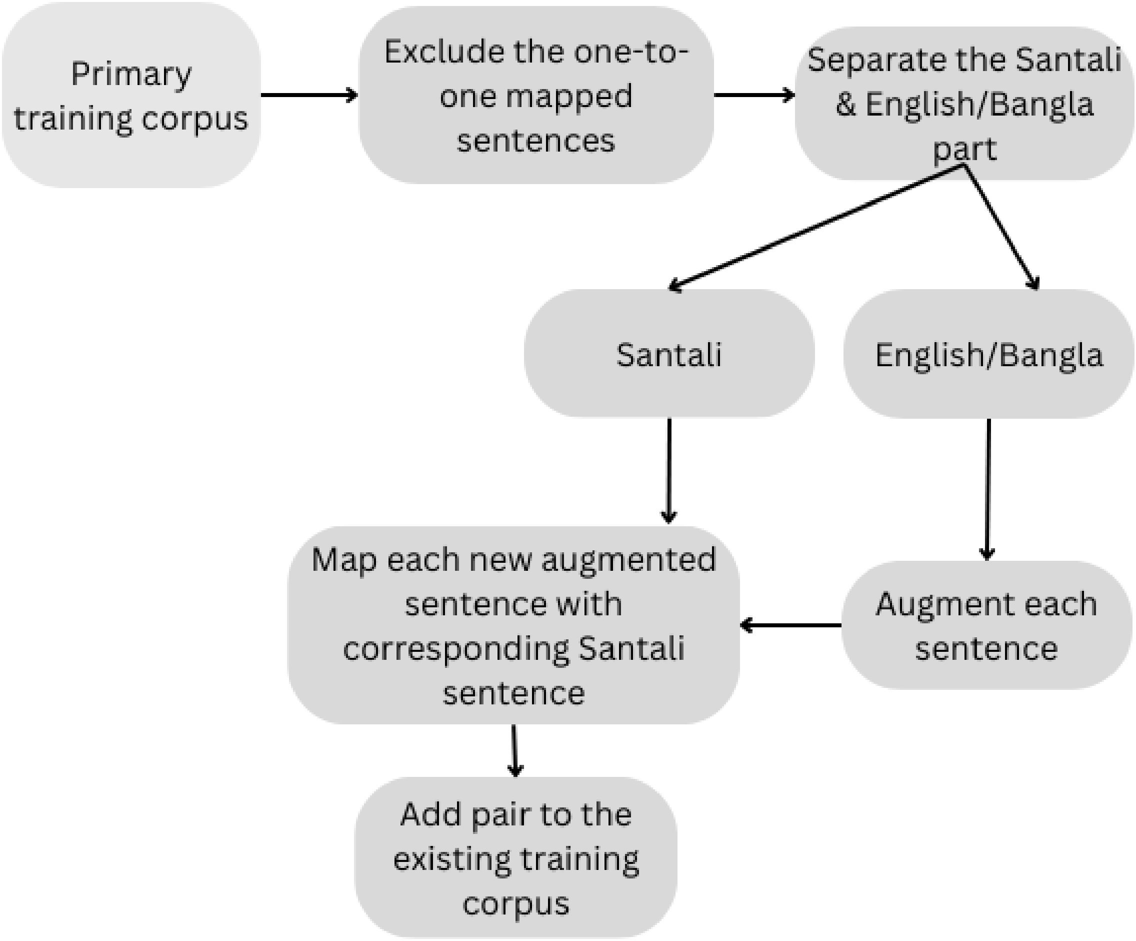

Primary training corpus
Exclude the one-to-one mapped sentences
Separate the Santali & English/Bangla part
Santali
English/Bangla
Map each new augmented sentence with corresponding Santali sentence
Augment each sentence
Add pair to the existing training corpus

Figure 5.5: Data Augmentation Flowchart

[1] https://github.com/sagorbrur/bnaug

5.3 Tokenizer

MT5 model is compatible with only sentence piece tokenizer.

For our Santali-English dataset, we extracted the Santali text from the dataset, and loaded the T5 tokenizer and trained it on out santali dataset, took the top 32,000 wordpieces as tokens which is here our source text. And for the english part, we used the t5-small tokenizer from Autotokenizer and tokenized out data.

After Tokenizing a sentence like - “re isor do ot ar sermae sirjaukeda”

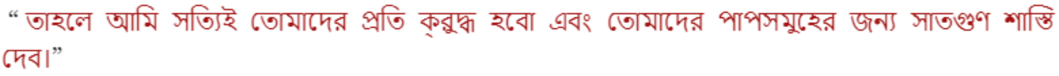
It becomes -

[“_re”, “_isor”, “_do”, “_ot”, “_ar”, “_sermae”, “_sirjau”, “_keda”]

Here, each wordpiece is assigned to a corresponding number.

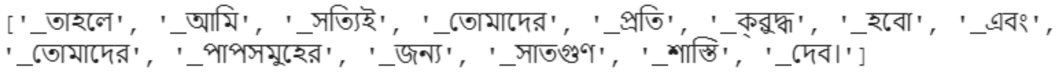
For the Santali-Bangla dataset, similarly first we extracted the Santali and Bangla text. We used the previosly trained Santali tokenizer to tokenized our source text. And for the Bangla text which is our target text, we used the pre-trained Bangla tokenizer [2].

From a sentence like this -



“তাহলে আমি সত্যিই তোমাদের প্রতি ক্রুদ্ধ হবো এবং তোমাদের পাপসমুহের জন্য সাতগুণ শাস্তি দেব।”

we get tokens-



```
['_তাহলে', '_আমি', '_সত্যিই', '_তোমাদের', '_প্রতি', '_ক্রুদ্ধ', '_হবো', '_এবং',
'_তোমাদের', '_পাপসমুহের', '_জন্য', '_সাতগুণ', '_শাস্তি', '_দেব।']
```

Here also all the tokens are mapped into a corresponding number.

5.4 Seq2Seq model

This model is implemented with two Recurrent Neural Networks. The first RNN is used to create the encoder part and using the second RNN we created the decoder part. The encoder part had A Embedding layer and then a GRU layer. In the encoder, the forward method takes in an input tensor and a hidden state tensor as input. Then the input tensor as embedded pass through the GRU, to obtain the output tensor, which gives the hidden state of the current time step. For the decoder, we created an Attention decoder RNN. The RNN consists of six layers- Embedding Layer, Attention layer, Attention Combination Layer, Dropout Layer, GRU layer, and Output Layer. We pass the hidden states, output language unique

[2] https://huggingface.co/csebuetnlp/banglat5

word size, and dropout probability to generate the next token. The forward method takes the input tensor, the hidden state tensor, and the encoder output tensor as the input. Then it embeds the input tensors and applies the attention mechanism to the encoder output tensor. Then in the GRU layer, the combined attention vector passes through and updates the hidden states. The attention mechanism helps the decoder to focus on different parts of the input sequence during the decoding process. The decoder return hidden states and output tensor and attention weights. We initialized the hidden size as 256 for the encoder and the attention decoder. We use the topk function to select the most probable next token to generate our sequence in a probability distribution. Abstract diagram of our model architecture.

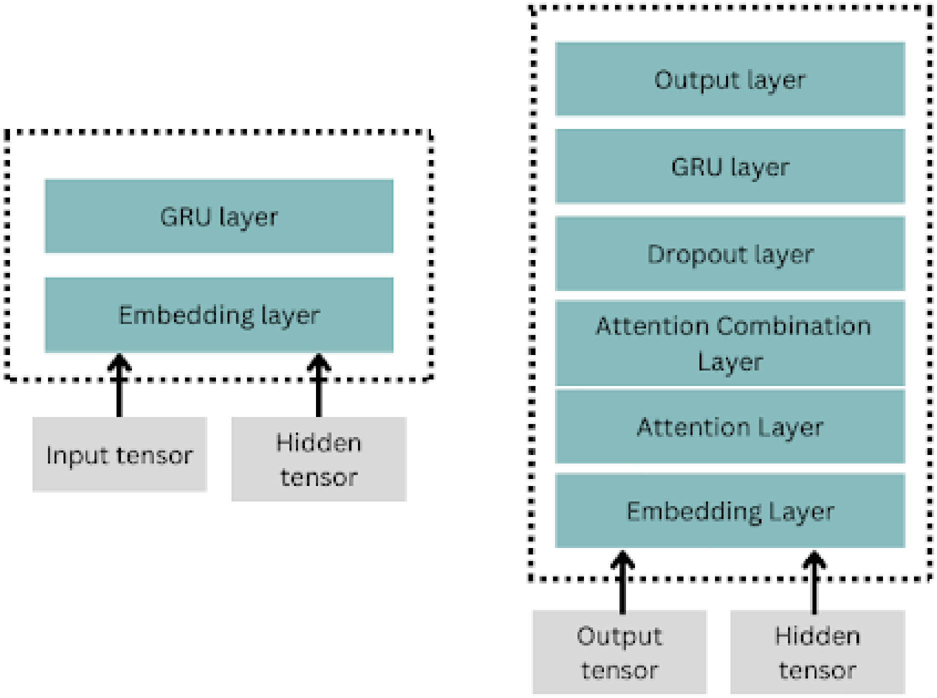


Figure 5.6: Encoder Decoder of Seq2Seq Model

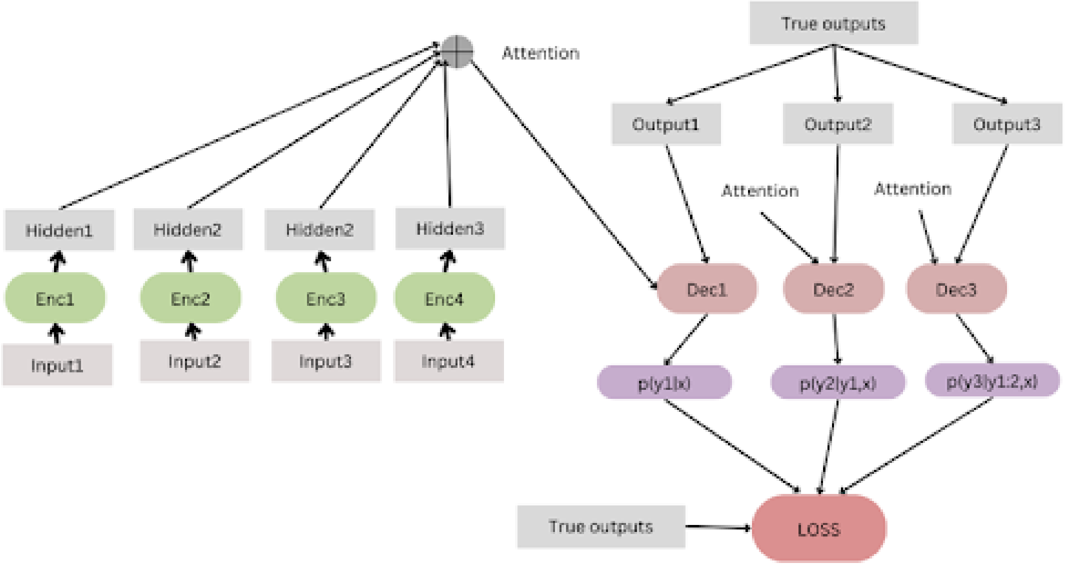


Figure 5.7: Seq2Seq Model with Attention

5.5 MT5 Architecture

In order to leverage transfer learning, we finetunes MT5-small checkpoints which was tranied on text and code of 101 different languages.It has only 300M parameters compared to 117 billion parameters of t5 model, making it a suitable choice to finetune on a smaller dataset.

The model follows the architecture of T5 [20] which is again a vanilla encoder-decoder model.

- **Encoder:**The encoder consists of 6 encoder layers, each with 12 attention heads. The attention heads are responsible for attending to different parts of the input sequence.
- **Decoder:**The decoder consists of 6 decoder layers, each with 12 attention heads. The decoder generates the output sequence one token at a time, by attending to the previous tokens in the sequence and the encoder output.

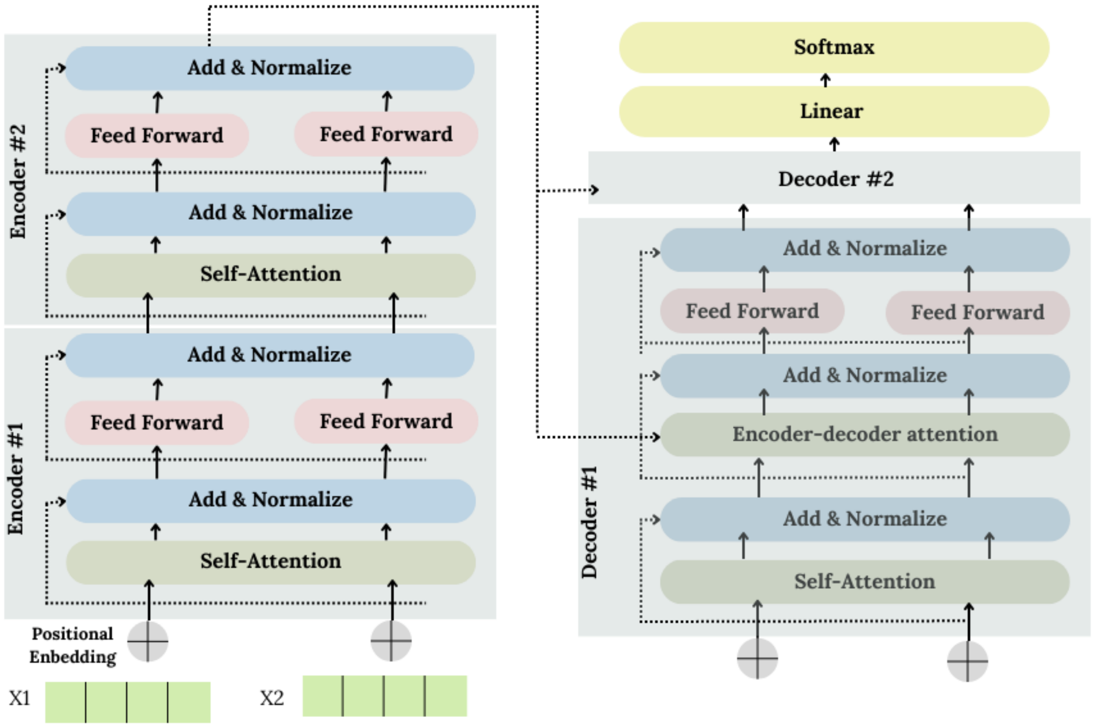


Figure 5.8: MT5 Model Architecture

T5 model is trained on multiple task such as -machine translation, classification task, sequence to sequence tasks like document summarization.
T5 model is trained with unsupervised objective so that the model can some way learn from the unlabeled data. It followed masked language modelling, where, some words are hidden form the original text and replaced with a token, the task of the model is to predict the hidden words.The model is taught to predict the tokens to find the texts that was dropped out. The objectives were-

1. randomly choose words in any position and replace them with any other words,then predict the original word
2. randomly shuffling the input test and try to predict the original text

Adding to the existing T5 model, in the MT5 model handles overfitting or underfitting of model due to sample ratio of low resource language and high resource language, with taking an approach by sampling examples according to the probability-

$$p(L) \propto |L|^{\alpha}$$

where $|L|$ is the number of examples in the language and $p(L)$ is the likelihood of selecting text from a specific language during pre-training. The hyperparameter α (typically with $\alpha < 1$) controls the boosting ratio of training on low-resource languages [**22**]. The paper showed $\alpha = 0.3$ to give a reasonable compromise between performance on high- and low-resource languages.

Chapter 6

Result Analysis

We have run two types of models. One is seq2seq and the other one is mT5 transformer. Seq2Seq is an industry standard model and is very good with sequential tasks as translation. But the limitation with this model is it needs ample amounts of data. Secondly we ran a transformer model named mT5. This model was previously pre-trained with many languages. But Santali was not one of them. We fine tuned this model with santali data to implement transfer learning. We used both augmented data and original data to see the difference.

6.1 Performance Evaluation

Model Type	**Validation score (BLEU)**	**Test score (BLEU)**
Seq2Seq	3.33	1.87
mT5 (Santali-English)	11.13	10.5
mT5 with augmentation (Santali-English)	8.93	6.79
mT5 with augmentation (Santali-Bangla)	2.85	1.59

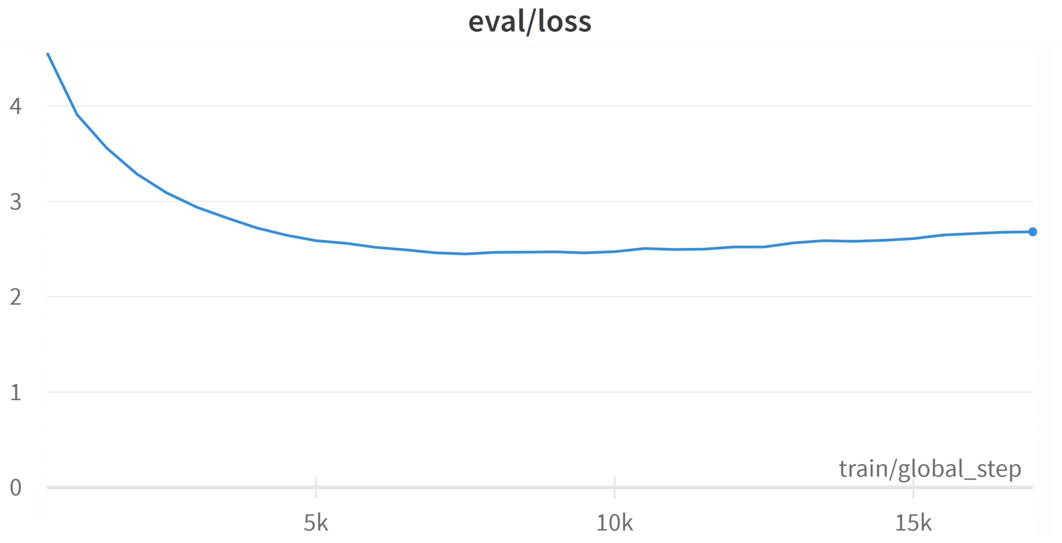


Figure 6.1: Santali-English Validation Loss (Without Augmentation)

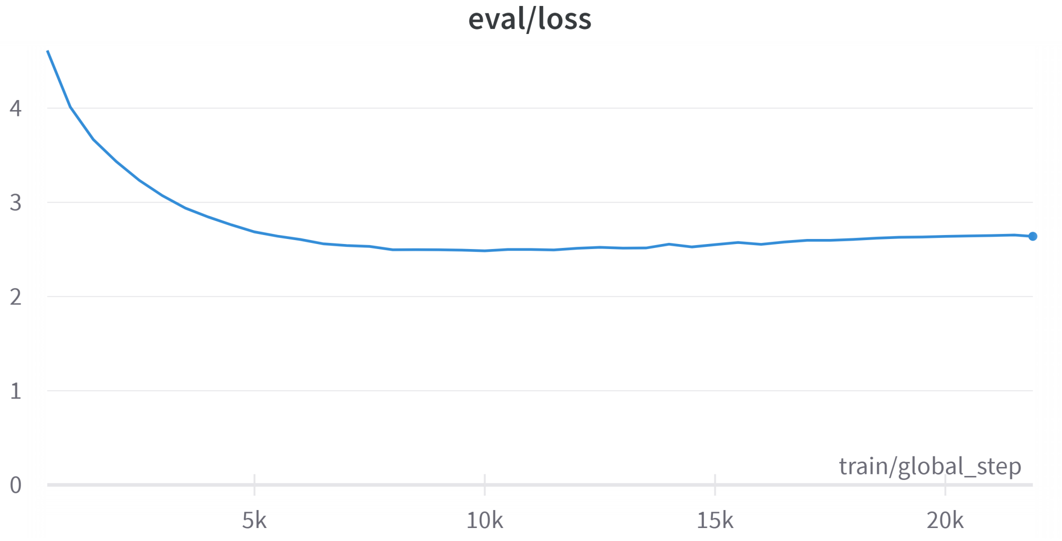
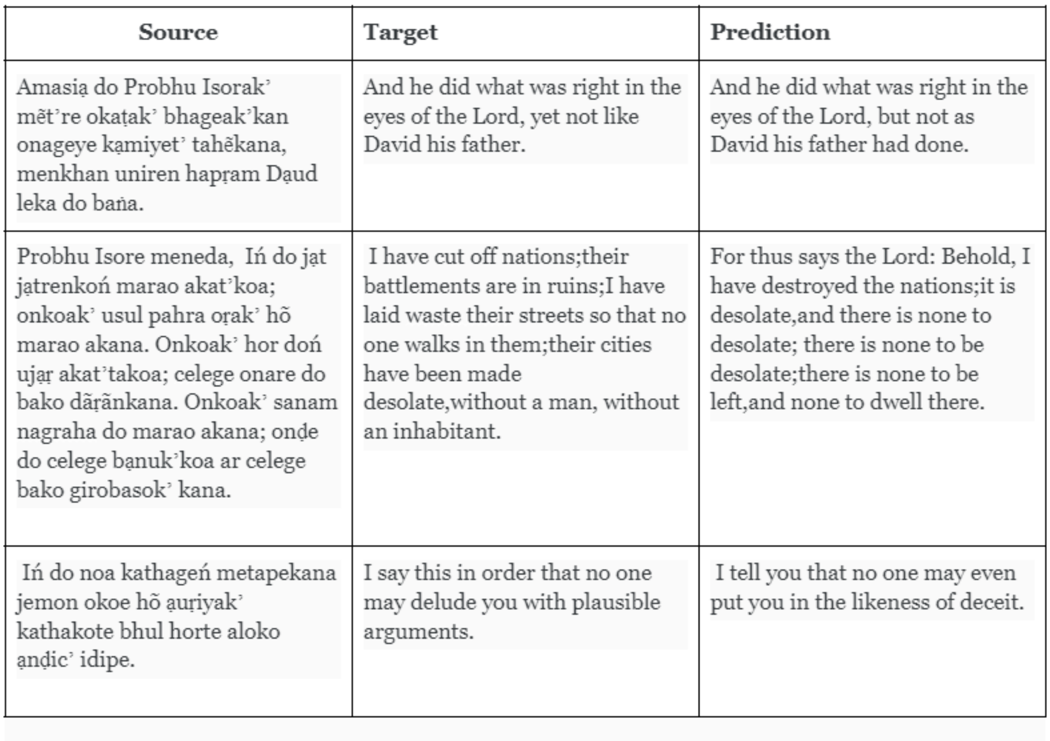


Figure 6.2: Santali-English Validation Loss (With Augmentation)

6.2 Model Predictions

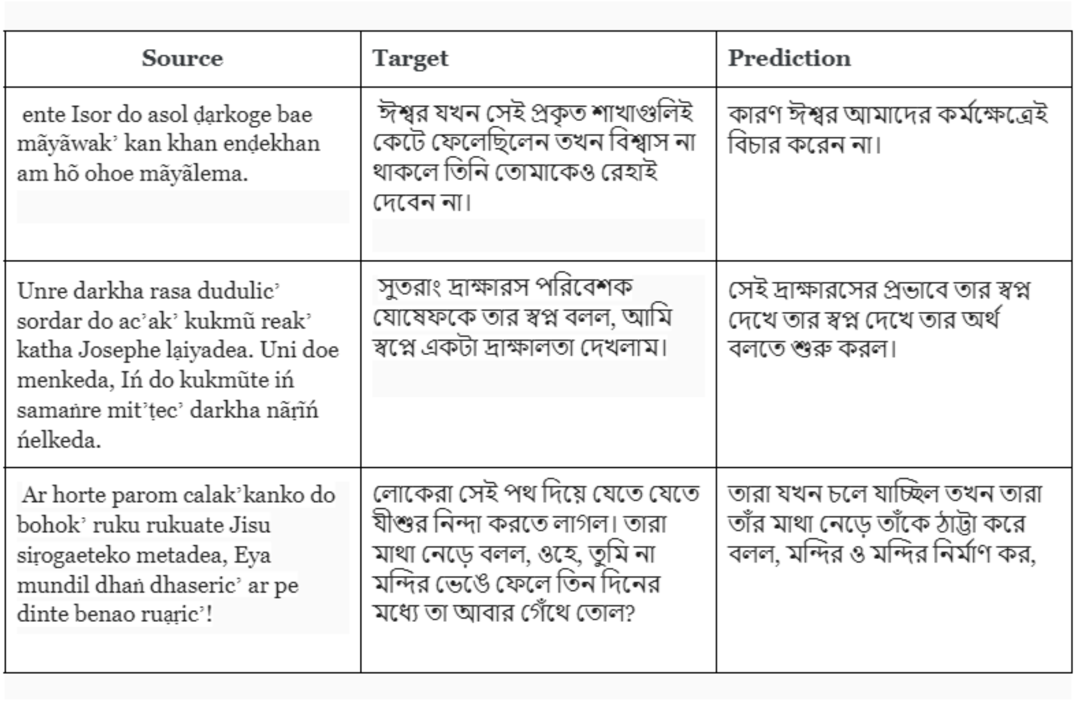
Santali-English with Augmented Data



Source	Target	Prediction
Amasiạ do Probhu Isorakʼ mẽtʼre okaṭakʼ bhageakʼkan onageye kạmiyetʼ tahẽkana, menkhan uniren hapṛam Dạud leka do baṅa.	And he did what was right in the eyes of the Lord, yet not like David his father.	And he did what was right in the eyes of the Lord, but not as David his father had done.
Probhu Isore meneda, Iń do jạt jạtrenkoń marao akatʼkoa; onkoakʼ usul pahra oṛakʼ hõ marao akana. Onkoakʼ hor doń ujạṛ akatʼtakoa; celege onare do bako dãṛãnkana. Onkoakʼ sanam nagraha do marao akana; onḍe do celege bạnukʼkoa ar celege bako girobasokʼ kana.	I have cut off nations;their battlements are in ruins;I have laid waste their streets so that no one walks in them;their cities have been made desolate,without a man, without an inhabitant.	For thus says the Lord: Behold, I have destroyed the nations;it is desolate,and there is none to desolate; there is none to be desolate;there is none to be left,and none to dwell there.
Iń do noa kathageń metapekana jemon okoe hõ ạuṛiyakʼ kathakote bhul horte aloko ạnḍicʼ idipe.	I say this in order that no one may delude you with plausible arguments.	I tell you that no one may even put you in the likeness of deceit.

Figure 6.3: Santali-English Model Predictions (With Augmentation)

Santali-Bangla with Augmented Data



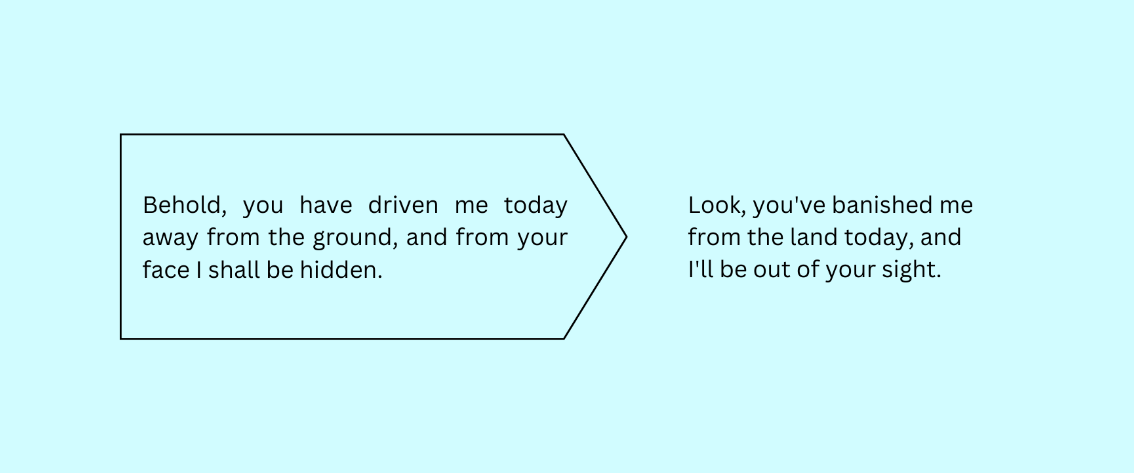
Source	Target	Prediction
ente Isor do asol ḍạrkoge bae mãyãwakʼ kan khan enḍekhan am hõ ohoe mãyãlema.	ঈশ্বর যখন সেই প্রকৃত শাখাগুলিই কেটে ফেলেছিলেন তখন বিশ্বাস না থাকলে তিনি তোমাকেও রেহাই দেবেন না।	কারণ ঈশ্বর আমাদের কর্মক্ষেত্রেই বিচার করেন না।
Unre darkha rasa dudulicʼ sordar do acʼakʼ kukmũ reakʼ katha Josephe lạiyadea. Uni doe menkeda, Iń do kukmũte iń samaṅre mitʼṭecʼ darkha nãṛĩń ńelkeda.	সুতরাং দ্রাক্ষারস পরিবেশক যোষেফকে তার স্বপ্ন বলল, আমি স্বপ্নে একটা দ্রাক্ষালতা দেখলাম।	সেই দ্রাক্ষারসের প্রভাবে তার স্বপ্ন দেখে তার স্বপ্ন দেখে তার অর্থ বলতে শুরু করল।
Ar horte parom calakʼkanko do bohokʼ ruku rukuate Jisu siṛogaeteko metadea, Eya mundil dhaṅ dhasericʼ ar pe dinte benao ruạṛicʼ!	লোকেরা সেই পথ দিয়ে যেতে যেতে যীশুর নিন্দা করতে লাগল। তারা মাথা নেড়ে বলল, ওহে, তুমি না মন্দির ভেঙে ফেলে তিন দিনের মধ্যে তা আবার গেঁথে তোল?	তারা যখন চলে যাচ্ছিল তখন তারা তাঁর মাথা নেড়ে তাঁকে ঠাট্টা করে বলল, মন্দির ও মন্দির নির্মাণ কর,

Figure 6.4: Santali-Bangla Model Predictions (With Augmentation)

6.3 Errors & Limitations

Variant alphabets:
Among the scarcity of Santali data, what intensifies the problem is the usage of different alphabet styles. Indian Santali has its own set of alphabets, and Bengali Santali has its own. As a result, for the few datasets that are available online, we cannot merge them.

Out-of-domain data:
Usually, transformer models need data from different domains during training to get rid of any biases. Since we trained with only Bible data, our prediction is that this model will not work with any conversational data. For example:



Behold, you have driven me today away from the ground, and from your face I shall be hidden.

Look, you've banished me from the land today, and I'll be out of your sight.

Figure 6.5: A verse from bible(left), how that verse would be used in today's conversations(right).

In the figure above, we can see a verse from the Bible in an arrowed box. We can easily see that the style of writing is not followed anymore. Since languages are not constant and, with time, we see significant changes in both conversational language and literature, our model may not successfully translate any modern-day literature. We could not verify this prediction yet due to the unavailability of other datasets.

Hardware limitations:
We started this project with a huge step-back due to its data lackings. Apart from that we had faced more adversities that hindered the process and outcome of our result. One of the main setbacks were hardware limitations. Any form of translation models are greatly data driven. Besides they need large memory spaces in RAM which we did not have. If we could use better machines, we could have used more advanced models.

Lack of Santali Expertise:
We did not have anyone who had knowledge of the Santali language. Moreover, the grammar of the Bengali-Santali writing style is not well documented. So we could not get good insights into the grammar of that language, which hindered us from creating complex augmented data. Lack of knowledge also stopped us from

creating a transliteration model from the Ol-Chiki script so that we could have used Indian-Santali data as well.

Chapter 7

Further WorkPlan

We intend to collect data on Santali-English parallel corpus to make an industry standard translation model. From wikipedia we have collected around 90k Santali sentences. However, this corpus is based upon OL-CHIKI script. To convert the OL-CHIKI script to our formatted script we have build a transliteration model. With help of some Santali experts, we sought to build a parallel corpus from this script.Our transliteration model can work as a bridge between the two domain of Santali scripts.

Additionally, we aim to collect more data by working directly with Santali community. So we will look for fund to start a crowd work project to collect data from local community and better hardware support. With these plans, we hope to generate a benchmark translation model for Santali in near future.

Chapter 8

Conclusion

In the era of globalization, the presence of translation models can elevate a particular language-speaking population to a global stage, diminishing the language barrier. Unfortunately, not all languages have their own translation model. Santali, a language spoken by seven million people across multiple countries in the world, is such a language. The main barrier to creating a Santali translation model is the shortage of data. But in the past few years, the field of NLP has crossed exemplary milestones and given models that can work even with very little data. Our primary objective was to determine the feasibility of constructing a translation model for the Santali language by exploring the limited online resources. Although the available internet resources were scarce, our adoption of advanced techniques such as transfer learning, augmented datasets, and transformer models yielded a commendable BLEU score. This success has instilled confidence in us that, with access to a substantial dataset, we can develop a practical Santali language translation model. Consequently, our goal is to collaborate with NGOs to gather data sets directly from the Santali community. Simultaneously, we plan to pursue financial support from government entities or private organizations to bolster our research efforts.

Bibliography

[1] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," *Computational Linguistics*, vol. 19, no. 1, pp. 75–102, 1993. [Online]. Available: https://aclanthology.org/J93-1004.

[2] R. C. Moore, "Fast and accurate sentence alignment of bilingual corpora," in *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Tiburon, USA: Springer, Oct. 2002, pp. 135–144. [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-45820-4_14.

[3] P. Halácsy, A. Kornai, V. Nagy, L. Németh, and V. Trón, "Parallel corpora for medium density languages," in Jan. 2007, pp. 247–258. DOI: 10.1075/cilt.292.32var.

[4] F. Braune and A. Fraser, "Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora," in *Coling 2010: Posters*, Beijing, China: Coling 2010 Organizing Committee, Aug. 2010, pp. 81–89. [Online]. Available: https://aclanthology.org/C10-2010.

[5] R. Sennrich and M. Volk, "Iterative, mt-based sentence alignment of parallel texts," May 2011. DOI: 10.5167/uzh-48036.

[6] S. Abdul-Rauf, M. Fishel, P. Lambert, S. Noubours, and R. Sennrich, "Extrinsic evaluation of sentence alignment systems," 2012.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ArXiv*, vol. 1409, Sep. 2014.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[9] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[10] B. K. U. M. A. R. Mohanta, *Invention of Alchiki Script for Santali Language: An Attempt to Preserve an Endangered Tribal Dialect.* 2015, ISBN: 9788183440691. [Online]. Available: https://www.academia.edu/14621750/Invention_of_Alchiki_Script_for_Santali_Language_An_attempt_to_Preserve_an_Endangered_Tribal_Dialect.

[11] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *arXiv preprint arXiv:1604.02201*, 2016.

[12] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," *arXiv preprint arXiv:1705.00440*, 2017.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] J. Gu, H. Hassan, J. Devlin, and V. O. Li, "Universal neural machine translation for extremely low resource languages," *arXiv preprint arXiv:1802.05368*, 2018.

[15] J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. Li, "Meta-learning for low-resource neural machine translation," *arXiv preprint arXiv:1808.08437*, 2018.

[16] A. Eftakhar, "Changing pattern of santali language: A trilingual situation and the emerging conflict," vol. 9, Jan. 2019, ISSN: 2519-5816.

[17] T. Hasan, A. Bhattacharjee, K. Samin, *et al.*, "Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation," *arXiv preprint arXiv:2009.09359*, 2020.

[18] S. Mia, "Describing phoneme inventory of santali language: A linguistic approach," Oct. 2020.

[19] J. X. Morris, E. Lifland, J. Y. Yoo, and Y. Qi, "Textattack: A framework for adversarial attacks in natural language processing," *CoRR*, vol. abs/2005.05909, 2020. arXiv: 2005.05909. [Online]. Available: https://arxiv.org/abs/2005.05909.

[20] C. Raffel, N. Shazeer, A. Roberts, *et al.*, *Exploring the limits of transfer learning with a unified text-to-text transformer*, 2020. arXiv: 1910.10683 [cs.LG].

[21] International Journal of Humanities, Social Sciences and Spirituality (IJHSSS), *Volume vii, issue v, september 2021*, 2021. DOI: DOI:10.29032. [Online]. Available: https://www.ijhsss.com/volume-vii,-issue-v,-september-2021.html.

[22] L. Xue, N. Constant, A. Roberts, *et al.*, "MT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. [Online]. Available: https://aclanthology.org/2021.naacl-main.41.

[23] P. R. Murmu, *A Brief History of Santali Language and Literature*. 2022, vol. 9, ISBN: ISBN:9788175250802, 8175250801.

[24] G. Loye. "Attention mechanism," FLOYDHUB. (2023), [Online]. Available: https://blog.floydhub.com/attention-mechanism/#luong-att.