A Translation-Based Approach to Sanskrit Text Summarization with the LongT5 Model

Parth Parikh
McDowell High School
parthparikh04@gmail.com

Abstract

Applying text summarization to Sanskrit can permit a heightened understanding of cultural and historical texts of South Asia and of other South Asian languages that are based on Sanskrit. Nonetheless, such an application does not currently seem to exist. This research specifically adopts a translation-based approach to text summarization to allow for the extension of natural language processing applications such as text summarization to low-resource languages such as Sanskrit that face data scarcity. When the quality of summarization was measured using the ROUGE scoring metric, a one-tailed paired t test consistently demonstrated p values below $\alpha = .05$, thus indicating a significant loss in quality for the translation-based approach in Sanskrit. The baseline adopted for summarization was direct summarization for each corresponding text in the English language.

1 Introduction

Since the emergence of text summarization in the 1950s, researchers have extensively worked to create increasingly human-like summaries through artificial intelligence. As defined by Merriam-Webster (2019), artificial intelligence is a branch of computer science simulating intelligent behavior in computers and encompasses numerous subfields including natural language processing (NLP), which is defined as the use of computers to process spoken languages. A subcategory of NLP, text summarization creates comprehensive and concise overviews of textual data (Syed et al., 2021).

As research in this field of text summarization has expanded, a wide majority of these applications have remained in English. Thus, most of the current state-of-the-art models for the task are trained on English datasets and are linguistically limited in their impact. With the translation-based approach outlined in this paper, the field of text summarization could rapidly expand to languages with little training data while simultaneously alleviating many expenses associated with model retraining.

This paper will begin by covering the overall need for a translation-based approach to text summarization through an analysis of previously conducted literature on the topic of text summarization, particularly that conducted for languages that face data scarcity, known as low-resource languages (LRLs). The paper will then detail the methodology adopted and analyze collected data before discussing the limitations of the study and possible directions for future work.

2 Background

As discussed in Section 1, substantial research has already been conducted on text summarization, but very few portions of this research consider the language of Sanskrit.

Figure 1: NLP Solutions and Population Sizes of Languages. Adapted from (Bhotia, 2022)

2.1 Languages

Increasing amounts of research in the field of NLP have emerged, they largely focused on high-resource languages (HRLs) that contain an abundance of resources for model training. Due to this distinction drawn between language adoption in research and resource availability, the majority of existing NLP research focuses on 20 HRLs out of over 7,000 total world languages (Chen and Verma, 2006; Roush, 2020; Guo et al., 2022). The remaining understudied languages have been termed LRLs, formally defined as languages less computerized due to data scarcity (Magueresse et al., 2020; Sinha and Jha, 2022). More specifically, these LRLs lack large text databases necessary for model training, leading to most preexisting research focusing primarily on HRLs such as English. As illustrated in Figure 1, this observation holds true despite the large population of speakers of LRLs, an incidence likely attributable to the significant allocation of time and funds necessary for the creation of robust databases for LRLs (Magueresse et al., 2020; Ranathunga et al., 2021).

To combat this discrepancy between resources available for LRLs and HRLs, researchers often choose to create new databases for LRLs; however, this approach requires more resources and time allocation than possible for the scope of this project (Magueresse et al., 2020). Namely, NLP models require large corpora of textual data for optimal training, and the creation of such vast databases entails a significant time commitment. For the language of Sanskrit, in particular, there are very few article-summary pairs available for training purposes. Moreover, the few that are present have not been publicly released (Sinha and Jha, 2022).

Despite resource scarcity surrounding Sanskrit, the language's semantics and content are incredibly rich. Linguistically, Sanskrit is viewed as a well-suited language for computers to understand due to the meticulous detail in the Aṣṭādhyāyī, the text outlining the language's grammatical principles. As a result of the aforementioned grammatical specificity, the language has minimal ambiguity in comparison to widely spoken languages such as English, a key benefit favoring its use in NLP (Bathulapalli et al., 2016; Bharati and Kulkarni, 2007). Beyond linguistics, Sanskrit texts detail key cultural components of South Asia, entailing works pertaining to both Hinduism and Buddhism (Hegarty, 2006; Pollock et al., 1990). However, the influence of Sanskrit extends far beyond the linguistic and religious realms: notably, scholars have discovered extensive detail in Sanskrit portrayals of secular historical events. Specifically, manual analyses of such texts have elicited a deeper understanding of Indian history before colonization, as demonstrated through previously unexamined connections between Jain and Mughal elites described in Sanskrit literature (Kaviraj, 2005; Truschke, 2012).

2.2 Extractive Text Summarization

Within the field of text summarization, one possible approach is known as extractive text summarization. This approach relies on selecting the most essential components of a given text, either in the form of complete sentences or as sentence fragments, and combining them to develop a summary of the inputted text (Allahyari et al., 2017). Covering a wide range of uses from competitive debate to medical texts, extractive summarization algorithms have been applied to various tasks (Roush, 2020; Chen and Verma, 2006). While this approach dominates the current scope of text summarization research for LRLs like Sanskrit, research in extractive summarization is still primarily concentrated on the English language (Tomar et al., 2021; Yadav et al., 2022). The few models that have been implemented in the past utilize the preexisting PageRank algorithm from English and fine-tune it to the language of Sanskrit (Das et al., 2016; Roush, 2020; Tomar et al., 2021). While some of these developments are relatively recent and reflect new algorithmic developments in extractive summarization, new state-of-the-art models have developed since their publication (Tomar et al., 2021).

With an extractive approach, issues arise primarily when analects are grammatically inaccurate due to the extraction of sentence fragments. Further, with query-based approaches, data can easily be misconstrued depending on the query sent to the model. For instance, biased wording within a given query would likely yield a summary not reflective of the original article's intentions (Roush, 2020). As Roush (2020) explains, biased queries such as "Economic decline causes unending war" would cause the model to extract text reinforcing the query rather than addressing a relatively unbiased input such as "Economic decline and war". In response, the focus of research has widely shifted to a newer approach to alleviate such concerns.

2.3 Abstractive Text Summarization

Namely, the novel approach designed to allay issues associated with extractive summarization is known as abstractive text summarization (ATS). ATS relies on a more human-like approach to text summarization that entails interpreting and processing the input document through NLP and generating new text through natural language generation (Rush et al., 2015; Sakhare and Kumar, 2016). Therefore, a summary generated through ATS would likely prove more grammatically and contextually sound than its extractive counterpart. The technique has also been extended to LRLs such as Chinese languages and the Indian language of Malayalam, thus representing an example of ATS in languages facing data scarcity similar to Sanskrit (Hu et al., 2016; Nambiar et al., 2021).

In a breakthrough from the previous architectures used for ATS, researchers at Google recently implemented the Transformer architecture in their LongT5 model. Partly attributable to the use of this novel approach, the LongT5 model reports Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores representative of a level of performance exceeding those of other popular models (Guo et al., 2022). However, likely owing to the lack of resources among LRLs as well as the size of the LongT5 model, there is minimal research available regarding the potential fine-tuning or retraining of this model on LRLs. As stated by Guo et al. (2022), the base and large models required 4x8 TPU v3 without model partitioning, and the xl model required 8x16 TPU v3 with 8 partitions. Hence, due to the vast allocation of resources necessary to retrain this model, doing so was not practically possible given the scope of budgetary allocation for this project.

Some researchers at the School of Sanskrit and Indic Studies pursued the creation of an ATS model for Sanskrit, their publication did not seem to report the creation of any conclusive model on account of the scarcity of Sanskrit training data (Sinha and Jha, 2020). Other publications relating to ATS in LRLs also did not seem to adopt the translation-based approach implemented by this research.

3 Method

This section will detail the precise study designed for this research. Due to the issue of data scarcity in Sanskrit, one of the main considerations in this regard began with the search for a suitable dataset.

3.1 Dataset

To analyze the effectiveness of text summarization with the LRL of Sanskrit via translation from English, this study first obtained a preexisting English database of texts. CNN and DailyMail news text and reference summary pairs were chosen for this study due to their relative recency and adoption by previous research (Hermann et al., 2015; see, 2017). In the database itself, the text and reference summaries were referred to as the article and highlights, respectively. While another possible approach would be the manual creation of a Sanskrit version of such a dataset, as noted in Section 2.1, the creation of any database, regardless of language, requires an extent of resource allocation beyond the scope of this project (Magueresse et al., 2020; Ranathunga et al., 2021). Moreover, the only existing Sanskrit database used by previous research is an extractive dataset of Wikipedia articles: a dataset that would produce suboptimal results since extractive summaries provide a poor baseline for abstractive summarization (Arora and Haptik, 2020). Specifically, some information from the source text will likely be omitted in the corresponding extractive reference summary, thus artificially deflating ROUGE scores reported. To create a database in Sanskrit, another potential source included the AIR Sanskrit News data. However, multiple issues were evident regarding this Sanskrit database. For one, the dataset occasionally entailed the summary being extracted directly from the full text, thus creating a poor baseline. Additionally, the conversion of this data to a processable format proved time-consuming; so, with this approach, testing would have been limited to a relatively small number of data samples.

On account of resource constraints with the computing system employed, rows of the CNN-DailyMail database were considered in groups of 500 text-summary pairs, a grouping that allowed for a reasonable runtime for each iteration and avoided excessive queries to the translation application programming interface (API).

3.2 Design

As demonstrated by the scripts corresponding with this research[1] and by the flowchart in Figure 2, the design for this research entailed a three-step process involving translation, summarization, and scoring. Upon iteratively loading sets of 500 text-summary pairs into a dictionary for ease of processing, each entry was validated to ensure no text or summary exceeded the 5,000-character limit imposed by Google Translate (Google, 2023b). Any summaries exceeding the aforementioned limit would prompt an error when using the translation tool; accordingly, the executed script removed any text-summary pairs that violated this criterion. Upon verification, the texts and summaries were translated through the translator module, which relies on Google Translate. The Google Neural Machine Translation model that powers Google Translate was the only available tool for English-to-Sanskrit text translation at the time of testing (Cloud, 2023; UlionTse, 2023). Thereafter, each English text was summarized with the LongT5 transient-global attention base-sized model implemented by Guo et al. (2022), with summaries appended to a list under the appropriate key of the data dictionary. After obtaining the generated summaries, the earlier process of translation through Google Translate was repeated for the generated summary.

Upon obtaining this data, similarity scores were calculated between the generated English and Sanskrit summaries and their corresponding references from the original database. ROUGE-1, ROUGE-2, and ROUGE-L scores were computed to measure similarity on each pair of summaries as these scoring metrics have frequently been used to measure the effectiveness of

[1] `https://github.com/ParthParikh04/Translation_Based_Sanskrit_TS`

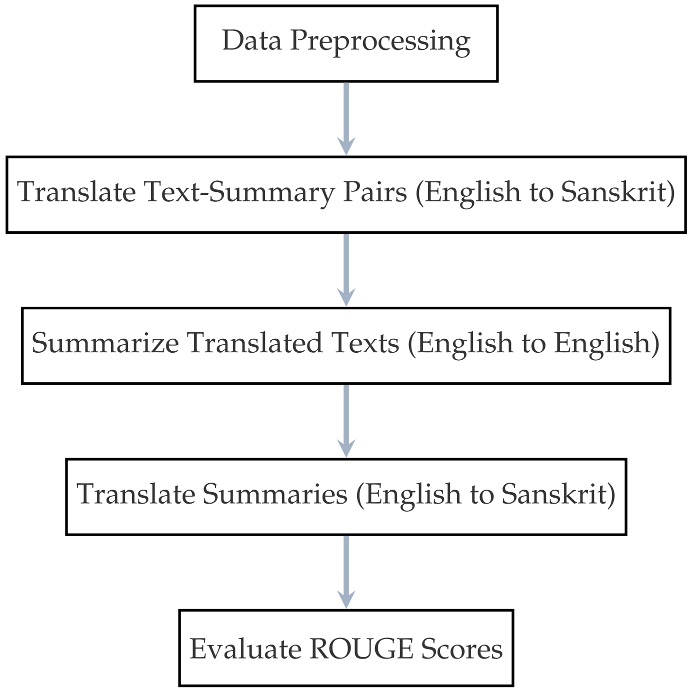

Figure 2: Design Flowchart

text summarization (Guo et al., 2022; Rush et al., 2015; Sakhare and Kumar, 2016). The ROUGE-1 and ROUGE-2 metrics are defined by considering 1-grams and 2-grams, respectively, where an n-gram is defined as a grouping of n words or tokens for an arbitrary positive integer n. The ROUGE-L metric, on the other hand, considers the longest common subsequence between the generated and reference summaries (Chiusano, 2022). For each of the aforementioned metrics, three scores were calculated:

$$\text{recall} = \frac{\text{number of n-grams in generated and reference summaries}}{\text{number of n-grams in reference summary}} \tag{1}$$

$$\text{precision} = \frac{\text{number of n-grams in generated and reference summaries}}{\text{number of n-grams in generated summary}} \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

For ROUGE score calculation, the now-archived ROUGE extension encompassed in Microsoft's GitHub repository[2] was utilized as it supported both Hindi and English scripts (Microsoft, 2023). Since Hindi and Sanskrit both use the Devanagari script, the module could also be used for calculating ROUGE scores on pairs of Sanskrit generated and reference summaries (Cardona, 2023). The Sanskrit texts processed by the ROUGE score calculator were thus submitted under the Hindi label.

Upon obtaining all ROUGE scores and storing them in separate files for each 500-sample batch, average ROUGE scores between all batches were evaluated.

3.3 Data Collection

Upon collection of data for as long as time permitted, 90,000 samples were tested; however, as discussed in Section 3.2, texts with 5,000 characters or more were not compatible with the Google Translate API for translation. Therefore, those samples were removed from consideration, thus leading to the ultimate calculation of ROUGE scores for 66,056 samples.

[2] `https://github.com/microsoft/nlp-recipes/blob/master/utils_nlp/eval/rouge/rouge_ext.py`

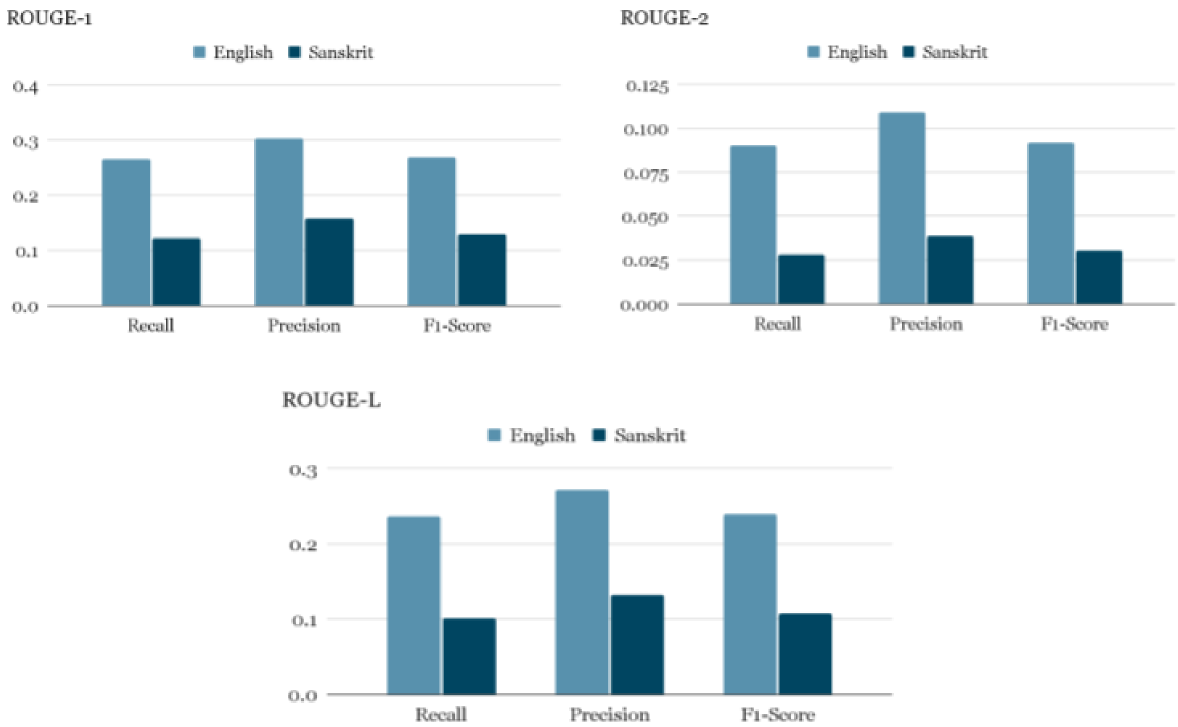

Figure 3: Reported ROUGE Scores Between Pairs of Generated and Reference Summaries

Also mentioned in Section 3.2, ROUGE-1, ROUGE-2, and ROUGE-L scores were computed between the generated and reference summaries. As per the pictorial depictions of these results in Figure 3, English summaries seemed to perform substantially better than their Sanskrit counterparts. Key extractions from this data are noteworthy and further support the previously noted observation: within the considered sample, 260 English results yielded all ROUGE scores of zeros, a condition met by 5,062 Sanskrit samples considered. These low scores correspond to near-zero common words between the generated and reference summaries, a scenario found to be more common in Sanskrit than in English.

3.4 Data Analysis

Between each pair of recall, precision, and F1 scores among ROUGE-1, ROUGE-2, and ROUGE-L metrics, a one-tailed paired t test was conducted. This test determined the statistical significance of the differences in ROUGE scores between English summaries and their respective Sanskrit counterparts. This test used an alpha level of .05, suggesting the results would be significant if $p < \alpha = .05$. With the paired t test, the p value represents the probability that, given each pair of English and Sanskrit summaries, the higher English ROUGE scores were purely by chance. The null hypothesis was stated to be no difference between English and Sanskrit ROUGE scores, while the alternate hypothesis was the English scores being higher than those for Sanskrit by a statistically significant margin. As the alternate hypothesis considered the English scores being specifically greater than Sanskrit scores rather than solely verifying a difference, a one-tailed test was most appropriate. In addition, a paired t test was fitting as scores were calculated for each pair of generated and reference summaries (Boston University School of Public Health, 2016; Kim, 2015).

A script presented in the GitHub repository linked in Section 3.2 calculated the reported p values. For proper execution of the aforementioned script, CSV files containing lists of ROUGE scores calculated on English summarization for each batch were consolidated into one spreadsheet with a row for each text-summary pair considered. A similar approach was taken for

	ROUGE Metric	English Average	Sanskrit Average	t value	p value	degrees of freedom
ROUGE-1	Recall	.27	.12	370.76	.001	66,055
	Precision	.30	.16	319.76	.001	66,055
	F1-Score	.27	.13	374.63	.001	66,055
ROUGE-2	Recall	.09	.03	195.05	.001	66,055
	Precision	.11	.04	176.72	.001	66,055
	F1-Score	.09	.03	196.59	.001	66,055
ROUGE-L	Recall	.24	.10	362.32	.001	66,055
	Precision	.27	.13	316.22	.001	66,055
	F1-Score	.24	.11	368.12	.001	66,055

Table 1: Average ROUGE Scores Reported Alongside Statistical Test Results

Sanskrit ROUGE scores in a separate file.

Upon observing the p values found in Table 1, as all $p < \alpha = .05$, the probability that ROUGE scores in English are greater than those in Sanskrit purely by chance is incredibly low, thus refuting the null hypothesis. Hence, ROUGE scores calculated on English summaries are statistically significantly higher than those on their corresponding Sanskrit counterparts. According to this analysis of p values, ROUGE scores for the Sanskrit translations of generated summaries created through the LongT5 model indicated significantly lower quality than those calculated for the original English summaries.

4 Limitations

As discussed in Section 3.4, the Sanskrit versions of the summaries from the LongT5 model yielded significantly lower scores relative to those presented between the English versions of generated and reference summaries. Consequently, translating text to Sanskrit via Google Translate is an ineffective method of ATS for the LRL of Sanskrit, thus refuting the initial hypothesis that reported English ROUGE scores would only be marginally higher than those reported for Sanskrit. Though such can be concluded, it cannot definitively be generalized to all LRLs or all summarization models due to the intrinsic limitations associated with this study, particularly those associated with the Sanskrit language and intermediate steps associated with the translation process.

4.1 Property-Oriented Language

When analyzing the difference between the linguistics of Sanskrit relative to other languages, such as English, one key consideration is the property-based nature of Sanskrit. According to the Monier-Williams Sanskrit-English Dictionary, when translating a word in English such as “water”, Sanskrit translations include multiple words including अंशूदक (aṃśūdaka), अकावङ्क (akāvaṅka), and अकुम्भ (akumbha) (Monier-Williams, 1960). Therefore, when comparing ROUGE scores between the Sanskrit and English samples, the translation tool could have mapped the same English word to different Sanskrit counterparts in different iterations. Meaning, in such scenarios, the ROUGE score formulas presented as formulas 1, 2, and 3 would yield artificially lower scores that do not appropriately reflect the quality of the text summarization process.

To elaborate, each Sanskrit word is based on the properties of the object it represents. For instance, when translating the English word "tree" through the Google Translate API, the Sanskrit result is वृक्षः (vṛkṣaḥ). The word वृक्षः, however, represents a single property associated with trees rather than the tree itself. Namely, it represents trees usually being cut or felled down. Stemming from this property, another word that can be used to represent a tree is the word पादप (pādapa), meaning "something that drinks using its feet" (gshah, 2011). In these scenarios, the number of directly similar words between the Sanskrit generated and reference summaries is artificially deflated, thus showcasing the potential for ROUGE scores yielding an inaccurate representation of true translation quality. Instead of the use of ROUGE scores, manual human scoring of the summaries is an alternative, albeit the costs and resource allocation of such an approach would prove to be major hindrances. This alternative approach will be further detailed in Section 5.

4.2 Omitted Texts

Beyond the realm of linguistics, technological limitations of Google Translate prevented larger text samples from undergoing the translation element of data collection. To allow for the creation of a complete Sanskrit database through translation, only elements of an English set of text-summary pairs that complied with the Google Translate API's 5,000-character limit were translated to Sanskrit and stored. On account of this constraint, during the execution of the script associated with this research, any text samples at or over the 5,000 character count limit were disregarded completely. Without this omission, scores reported on summaries of longer text samples between English and Sanskrit may have proved to differ from those reported, although the extent and direction of this difference are unclear.

4.3 Issues in Translation

When translating samples through the Google Translate API, another issue lay in incomplete translations. Within the example depicted below, not all parts of the generated summary in English are translated when producing the Sanskrit version:

> Generated Summary in English: "The former Los Angeles Lakers player said when he was first told he had HIV he was convinced he was going to die, but advances in drugs has helped Johnson - and millions of others - survive. Johnson, who became the face of HIV/Aids 22 years ago, is now campaigning for more people to get tested for the disease, especially those in black or Hispanic communities. Johnson said that Glaser, whose HIV had developed to Aids, was able to answer questions from him and his wife Cookie, who was two months' pregnant at the time, about living with the disease."
>
> Translated Generated Summary: "पूर्वः लॉस एन्जल्स-लेकर्स्-क्रीडकः अवदत् यदा प्रथमवारं तस्मै एच.आई.वी. जॉन्सन्, यः २२ वर्षपूर्वं एच.आई.वी. जॉन्सन् उक्तवान् यत् यस्य एच.आई.वी."

Upon deeper analysis of this issue, omission of the "/" symbol in the phrase "...who became the face of HIV/AIDS 22 years ago" would allow the script to create a full translation; however, omission of such characters in other scenarios elicits unforeseen issues. Namely, when removing other non-alphanumeric characters, such as the "£" symbol, the phrase "and 'contaminated' the £500 chairs" became "and 'contaminated' the 500 chairs", thus drastically changing the meaning of the sentence. After considering this dilemma that arises from the removal of certain characters to rectify the issue of incomplete translations, this approach of removing non-alphanumeric characters was not adopted. Nonetheless, if such issues had not been present in the Google Translate API, Sanskrit ROUGE scores would almost certainly have been higher than those computed by this study. As a result of this translation issue, some Sanskrit scores were calculated to be zero for text samples where the English scores were non-zero, suggesting a loss in accuracy when translating the generated summary to Sanskrit. In part, the low p value reported in Section 3.4 was likely a result of this issue in the translation process as it would artificially deflate Sanskrit ROUGE averages; however, the extent of this impact remains uncertain.

5 Future Work

Considering the low reported Sanskrit scores when summarizing texts through translation, future research can take multiple avenues. Even so, to pursue most of these avenues of research in LRLs, future researchers must create new databases to reduce the issue of data scarcity that prevents the research of large NLP models on LRLs. While this will undoubtedly prove expensive, it is a necessary step to further research in languages facing data scarcity. Such an approach would lessen many of the limitations of this study, namely those corresponding to issues in the Google Translate API, and would allow for additional research into the primary implication of this study, which will be discussed in Section 6.

5.1 Human Scoring

With the currently adopted methodology, recall that the use of ROUGE scores to measure the performance of text summarization led to limitations when considering linguistic differences between Sanskrit and English and issues in translation. These issues prompt the possibility of an alternative scoring approach; namely, a more comprehensive score can be assigned by human scorers to the generated summaries, as previously proposed by Allahyari et al. (2017). Through such an approach, scoring would not only limit issues arising from linguistic differences between English and Sanskrit but also would consider aspects disregarded by ROUGE scores, such as sentence structure and grammar (Allahyari et al., 2017). To implement such a manual scoring approach, professional scorers could be instructed to rate each Sanskrit and English summary on a Likert scale, which uses a numerical rating approach, in terms of retained meaning from the source text and grammatical fluency. Thereafter, a one-tailed paired t test can be executed between each pair of scores assigned to Sanskrit and English summaries.

While this approach would account for factors not considered by ROUGE scores, it was not deemed appropriate for this particular study for a multitude of reasons. For one, the resource constraints associated with gathering a group of willing participants with sufficient knowledge of Sanskrit and English were largely unfeasible given the scope of this research. Further, manual scoring of these summaries by the sole researcher was impracticable due to time constraints and possible sources of bias this could elicit. Beyond resource constraints, adopting a human scoring approach could lead to increased subjectivity unless a formula for scoring summaries that accounted for grammatical fluency was created, a task once again beyond the scope of this project. Accordingly, while a human scoring approach remains a suitable avenue for future research with fewer time and resource constraints, it was not adopted for this particular study.

5.2 Other NLP Tasks

Bringing another avenue for future research to light, the broad field of NLP encompasses many tasks beyond solely text summarization. Notably, corporations frequently use the NLP task of semantic analysis to understand reviews without wasting employee time and resources on this task. Nonetheless, reviewers and clients better acquainted with LRLs like Sanskrit are still unable to benefit from such applications due to the limited research in Sanskrit semantic analysis. By expanding such research to LRLs, corporations would be able to address the needs of a wider audience, unlike the restricted audience the current state of research considers.

Further, predictive text, a more common use of NLP that has recently expanded to applications created by Google and Apple, can also be expanded to allow for the use of artificial intelligence to build on writing in LRLs (Inc., 2023; Google, 2023a; Tableau, 2023). The uses of such NLP tasks are often as pervasive as those for text summarization and are infrequently researched for LRLs due to the issue of data scarcity. So, if future research is to expand such applications to Sanskrit, more individuals can enjoy the benefits thereof. In a more generalized sense, by expanding NLP applications to LRLs, modern technological implementations can be expanded to populations beyond those of the Western world.

5.3 Other LRLs

To discern whether the substantial loss in quality when conducting text summarization through a translation intermediary is limited to Sanskrit, future research can analyze the extent of such loss in quality for other LRLs through ATS. To pursue such research, one of the first steps would entail creating a reliable calculation of ROUGE scores for that language. As ROUGE scores' calculations are language-specific, the existing resources used to calculate the scores will likely require some degree of manipulation before such a research project is properly executed. This avenue for research retains much potential despite the results of this study as it may present a translation-based approach to achieving NLP tasks as a viable approach for some other LRLs. Such findings could present a novel alternative to the laborious and resource-intensive task of model retraining. Such research on non-Sanskrit LRLs will also allow for a deeper understanding of the linguistics of Sanskrit in comparison to properties of other languages, primarily in the contexts of their influences on ROUGE scores, thus investigating a key limitation of this research.

In general, as reported in Section 2, the scope of research with LRLs is slim, especially in NLP, thus building upon the necessity for future research in this field. Increasing research on NLP applications such as text summarization with other LRLs will allow for a deeper historical and cultural understanding from lesser-known perspectives that are often difficult to understand due to the dwindling count of knowledgeable speakers of the languages in which they are scripted.

6 Implications

Considering the findings of this research alongside the inefficiency of summarizing Sanskrit texts through translation from English, perhaps the most significant future research would be to retrain a large language model such as the LongT5 model on Sanskrit texts. As an implication of this paper's findings, once considering the results of this study as well as the limitations thereof, retraining the LongT5 model on the Sanskrit language would likely abate the linguistic and technological issues noted earlier. In addition, by calculating ROUGE scores of summaries produced while testing this retrained model, one would be able to test for the statistical significance of the difference between ROUGE scores when using the method proposed in this paper and those produced with a retrained model. As the intermediate step of translation is suspected to reduce the quality of the summary and detract from the overall meaning of the source text, one can hypothesize that this retraining approach will prove more effective in producing high-quality summaries. Researchers can then apply such summarization models to better understand previously-unread Sanskrit texts.

Adopting such an approach would not only require the creation of a database of Sanskrit texts and high-quality abstractive summaries but also would require the technological and financial resources to load a large text summarization model for retraining. Future research should thus create high-quality text-summary pairs in Sanskrit and allocate necessary resources to retrain state-of-the-art models such as the LongT5 model. By doing so, researchers can preserve the richness of Sanskrit texts and the cultural values therein. In fact, by increasingly researching modern artificially intelligent applications in Sanskrit, societal understanding of Indian history can be augmented, as has previously been seen with manual analyses of Sanskrit texts (Kaviraj, 2005; Truschke, 2012). More broadly, by better understanding the language foundational to other Indian languages, researchers can better understand languages spoken across the South Asian subcontinent and texts in those languages (Ministry of Human Resource Development, 2020; Gupta and Pradhan, 2021). Especially considering the extensive population of speakers of Indian languages, the benefits of increasing the study of Sanskrit would be far-reaching in their ability to encourage social progress. Even from scientific and mathematical perspectives, ancient Sanskrit texts list foundational properties and ideas not formally documented until much later, thus suggesting the possibility of new discoveries in those fields by increasing NLP research for Sanskrit (Chopra, 2020; Nahata, 2021).

By understanding Sanskrit in more detail and dedicating more resources to the development of NLP applications therein, one can expect extensive impacts that better South Asian society and augment current understandings of mathematics and the sciences.

7 Acknowledgements

This research would never have been possible without the support of my research advisor, Mr. Robert Hodgson, who has provided guidance and encouragement to me throughout this research process; Dr. Avinash Varna, who has provided technical mentorship during the implementation of the methodology; the administration at McDowell High School who has supported the AP Research program; the Sanskrit as a Foreign Language program for kindling my love for Sanskrit; and all who have reviewed my paper and provided suggestions without which this research would never have reached its current state. I am sincerely grateful to all who have contributed to making this paper possible.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text summarization techniques: A brief survey. arXiv:1707.02268 [cs], 07.

Gaurav Arora and Jio Haptik. 2020. inltk: Natural language toolkit for indic languages.

Chandana Bathulapalli, Drumil Desai, and Manasi Kanhere. 2016. Use of sanskrit for natural language processing. 78 International Journal of Sanskrit Research, 2:78–81.

Akshar Bharati and Amba Kulkarni. 2007. Sanskrit and computational linguistics.

Tenzin Singhay Bhotia. 2022. Challenges in using nlp for low-resource languages and how neuralspace solves them. 01.

Boston University School of Public Health. 2016. Paired t-test.

George Cardona. 2023. Devanāgarī | writing system. 02.

Ping Chen and R. Verma. 2006. A query-based medical information summarization system using ontology knowledge. page 37–42, 06.

Fabio Chiusano. 2022. Two minutes nlp — learn the rouge metric by examples. 04.

Anil Chopra. 2020. Vimana: The ancient indian aerospace craft – time for indigenisation. 08.

Google Cloud. 2023. Automl translation beginner's guide | automl translation documentation, 03.

Swagatam Das, Tandra Pal, Samarjit Kar, Suresh Chandra Satapathy, and Jyotsna Kumar Mandal, editors. 2016. Query-based extractive text summarization for sanskrit. Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015, Springer India.

Google. 2023a. Manage google autocomplete predictions - android - google search help.

Google. 2023b. Translate written words.

gshah. 2011. Features of sanskrit that make it an extra-ordinary language. 04.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. arXiv:2112.07916 [cs], 05.

Mohini Gupta and Uma Pradhan. 2021. Can sanskrit education be inclusive?: A perspective on the new education policy 2020. 06.

James M. Hegarty. 2006. Encompassing the sacrifice: On the narrative construction of the significant past in the sanskrit mahābhārata. Acta Orientalia Vilnensia, 7:77–118, 01.

Karl Moritz Hermann, Tomáš Ko□iský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. arXiv:1506.03340 [cs], 11.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2016. Lcsts: A large scale chinese short text summarization dataset. 02.

Apple Inc. 2023. Use predictive text on iphone.

Sudipta Kaviraj. 2005. The sudden death of sanskrit knowledge. Journal of Indian Philosophy, 33:119–142, 02.

Tae Kyun Kim. 2015. T test as a parametric statistic. Korean Journal of Anesthesiology, 68:540–546.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. arXiv:2006.07264 [cs], 06.

Merriam-Webster. 2019. Definition of artificial intelligence.

Microsoft. 2023. Nlp best practices, 03.

Ministry of Human Resource Development. 2020. National education policy 2020 ministry of human resource development government of india. Technical report.

Sir Monier-Williams, Monier. 1960. A Sanskrit-English dictionary, etymologically and philologically arranged, with special reference to cognate Indo-European languages. new ed., greatly enl. and improved, with the collaboration of E. Leumann, C. Cappeller and other scholars. Oxford Clarendon Press.

Shree Nahata. 2021. The importance of sanskrit in indian education. 06.

Sindhya K Nambiar, David Peter, and Sumam Mary Idicula. 2021. Abstractive summarization of malayalam document using sequence to sequence model. 1:347–352.

Sheldon Pollock, Pamela Stewart, and Andrew Strathern. 1990. From discourse of ritual to discourse of power in sanskrit culture. Source: Journal of Ritual Studies, 4:315–345.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey. arXiv:2106.15115 [cs], 06.

Allen Roush. 2020. Cx db8: A queryable extractive summarizer and semantic search engine. arXiv:2012.03942 [cs], 12.

Alexander Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for sentence summarization. pages 17–21. Association for Computational Linguistics.

D.Y. Sakhare and Raj Kumar. 2016. Syntactical knowledge and sanskrit memamsa principle based hybrid approach for text summarization. International Journal of Computer Science and Information Security (IJCSIS), 14, 04.

2017. Get to the point: Summarization with pointer-generator networks. Association for Computational Linguistics, 07.

Shagun Sinha and Girish Jha. 2020. Abstractive text summarization for sanskrit prose: A study of methods and approaches. page 60–65, 05.

Shagun Sinha and Girish Nath Jha. 2022. An overview of indian language datasets used for text summarization. arXiv:2203.16127 [cs], 04.

Ayesha Ayub Syed, Ford Lumban Gaol, and Tokuro Matsuo. 2021. A survey of the state-of-the-art models in neural abstractive text summarization. IEEE Access, 9:13248–13265.

Tableau. 2023. 8 natural language processing (nlp) examples.

Shalini Tomar, Sandeep Rana, and Baldivya Mitra. 2021. Sanskrit language text summarization using modified page rank algorithm. JOURNAL OF XI'AN UNIVERSITY OF ARCHITECTURE & TECHNOLOGY, XIII, 06.

Audrey Truschke. 2012. Setting the record wrong: a sanskrit vision of mughal conquests. South Asian History and Culture, 3:373–396, 07.

UlionTse, 2023. Translators. PyPI, 01.

Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. 2022. Automatic text summarization methods: A comprehensive review. arXiv:2204.01849 [cs], 03.