INFORMATICS INSTITUTE OF TECHNOLOGY
In Collaboration with
UNIVERSITY OF WESTMINSTER

SinDOC:
A Combined Approach of Summarizing Low Resource Sinhala Language Documents

A dissertation by
Mr. Mohamed Hamza Ziyard

Supervised by
Mr. Mithushan Jalangan

Submitted in partial fulfillment of the requirements for the BSc. (Hons) Computer Science degree at the University of Westminster.

May 2023

ABSTRACT

Summarization is one of the NLP related tasks that is widely researched during the recent times. Due to the large amount of data stored on the internet, users do not consume all information that is available. They will be looking for a small piece of information that will convey the most important information to them as quickly as possible. Sinhala language is one of the low resource languages which do not have many contributions within the field of document summarization. Mainly because of the lack of resources. To address this problem the author has come up with a novel approach that uses a combined method to summarize Sinhala documents by using both extractive and abstractive techniques.

The proposed model uses word frequency and sentence scoring approaches for the extractive model and uses a pre-trained model for the abstractive summarization model. The author has presented a new dataset that has been translated from an existing English dataset. The pre-trained model uses this dataset for training. The author was also able to prove that automating hyper parameter tuning to generate training arguments for the abstractive model gives better results with less time constraints when compared to the traditional approaches. The author has also given the option of generating summaries through all three approaches to make sure that the user gets the best overall result. The model build in this research shows good results that outperforms previous works.

Overall, this research proposes a combined approach of generating summaries for Sinhala documents. As all the models and code are made publicly available the author believes that this work could build a strong foundation for future researchers.

Keywords: Natural Language Processing, Extractive summarization, Abstractive summarization, Combined summarization, Sinhala Language

Subject Descriptors:

- Computing methodologies → Artificial Intelligence → Natural language processing → Information extraction
- Theory of computation → Theory and algorithms for application domains → Machine learning theory → Semi-supervised learning
- Human centered computing → Interaction design → Interaction design process and methods → User centered design