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

Mohamed Hamza Ziyard
Department of Computing
Informatics Institute of Technology
Colombo,Sri Lanka
hamza.2019407@iit.ac.lk

Mithushan Jalangan
Department of Computing
Informatics Institute of Technology
Colombo,Sri Lanka
mithushancj@gmail.com

***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 model build in this research shows good results that outperforms previous works.**

***Index Terms*—Natural Language Processing, Extractive summarization, Abstractive summarization, Combined summarization, Sinhala Language**

I. Introduction

This research will mainly focus on the Sri Lankan community who will have the difficulty in summarizing long Sinhala documents and getting the idea of it in simple terms. Sinhala language is the native language of all Sri Lankans. Many native language speakers and readers tend to miss the most important information in a document as they get bored or stop continuing to read a long document. This research focuses on providing summaries for such long Sinhala documents and help in conveying the most important context of a document to the people. Details of existing work, author's working schedules, project management methodologies to be used for this research are stated in the upcoming sections.

II. Problem Domain

A. Natural Language Processing

Natural language processing is the process in which computers are trained for text and spoken related tasks that provide results as humans. [1]. During recent times NLP has been a highly talked about and contributed research area. Some of the applications that are widely useful are machine translation, document translation, question answering and text generation [1]. Industries such as healthcare, finance and customer service are highly benefited from NLP's use cases.

B. Low resource languages

English related projects have been growing widely in the field of Natural Language Processing due to the extensive amount of data available in any type of platform. But languages like Sinhala are languages that do not have a wide resource available on the internet. These types of documents are known as low resource languages [2]. Even though there isn't much data on the internet, such a language is considered to be the primary language of a country. One such use case would be that there are around 21 million people within Sri Lanka who are considered to have Sinhala as their primary language and around 11.5 million internet users [3], which means that even though it is a low resource language around the world it still has a lot of value within Sri Lanka. Summarization is one of the tasks that haven't been much focused when it comes to low resource languages [2]. The author will be solving this problem by implementing a summarization solution for the low resource language Sinhala.

C. Document Summarization

Document summarization is the task of collecting data from a long document content and summarizing it into shorter simpler sentences while maintaining the main content of the document [4]. Document summarization is mainly of two categories namely extractive and abstractive summarization. Extractive summarization is a methodology of extracting the most important sentences or words from a document and generates a summary while abstractive summarization uses the technique of paraphrasing the documents content while generating a summary [5]. Hybrid approaches include both extractive and abstractive techniques for summarization [6]. During recent times researchers have been actively contributing on single document and multiple document summarization using both extractive and abstractive techniques [7].

III. Litreature Review

A. NLP for Sinhala Language

Dhananjaya et al. (2022) [8] proposes a new model called SinBERT for text classification tasks. The model has been created by transfer learning RoBERTa [9]. Additionally, the authors of this research have introduced two datasets that are specific for Sinhala text classification tasks [8]. Compared to

Identify applicable funding agency here. If none, delete this.

previous state-of-the-art methods their model performs best for classification tasks due to their larger number datasets with unique values [8]. The model has been trained with four main classification data namely sentiment analysis, writing style classification, news source classification and news categorization whereas writing style and news source classification datasets were introduced by the authors [8]. SinBERT-small and SinBERT-large were introduced in this paper however the authors recommend using SinBERT-small due to the less accuracy of the large model. They have stated that the issue of less accuracy is because of the less amount of data available for Sinhala that needs to be passed into a higher number of hidden layers within the model [8]. The authors are interested in testing the models for other NLP tasks and improve the model by feeding more data into it. Rathnasena et al. (2018) [10], proposes a method of using computer vision and summarization methods to store and search through old Sinhala books. The method includes Optical Character Recognition which extracts test information through images and then uses extractive summarization techniques to extract the most important information of the document. Due to the improvement of technology of storing old preserves the author identifies that there are no such methods for storing old sinhala books. Due to that the author decides to build a model that automates the process of recognizing characters from old Sinhala books and generating summaries of them to store for future readings [10]. The lack of sinhala corpus and the smaller number of historical words stored tend to reduce the accuracy of the model by some slight percentage. Additionally, the author also recommends abstractive summarization techniques for future work because of the paraphrasing and restructuring it could provide for a document [10]. The use of deep learning techniques and RNN's are recommended for future works [10].

B. *Extractive summarization*

Deshpande and Jahirabadkar (2021) [11] research focus on extractive summarization methodologies done using lexical chaining approach and pre-trained model BERT [12]. The model proposed is trained using Devanagi script data which includes Hindi Language [11]. The author has tested the lexical approach initially and then trained the dataset for the BERT models. The author provides some details on the outputs but does not include any test results, so it is hard to confirm if it performs better than the SOTA approaches. The author highlights that low resource languages are an important domain that should be considered for text summarization [11].
Zhang, Wei and Zhou (2019) [13] proposes a method of using a pre-trained model for better extractive summarization techniques.The model is fine-tuned on top of BERT [12]. The inclusion of hierarchical transforms to BERT implements their model HIBERT [13]. The model successfully surpasses the SOTA models as demonstrated using the necessary evaluation methods. The authors state that the model could be improved with a larger dataset [13]. They would also like to take the model to the extent of having the possibility of including a question answer system within the documents [13]. Additionally, they would also like to improve the architecture to provide better results.
Sakhadeo and Srivastava, (2018) [14] propose a method of extractive summarization that includes word frequency base approaches and entity graph generation. The authors have conducted both qualitative and quantitative approaches of testing to evaluate the performance of their model. They were able to identify that the quantitative results show that their model outperformed the SOTA methods while their human evaluation or qualitative methods show that most of the human did highly prefer their models' generated summaries when compared to previous SOTA models [14]. The authors were able to solve limitations of both approaches word frequency and entity relationship generation by creating a hybrid model that includes both the approaches [14]. Identifying a suitable normalized function and improving the model for multilingual tasks are future improvements that the authors have stated.

C. *Abstractive summarization*

Hasan et al. (2021) [15] is a work on abstract summarization tasks that introduces a multilingual dataset of 44 languages. The research was mainly focused on low-resource languages and its aim was to build a model that can be used for multiple languages in the field of summarization. It is also stated to have the largest abstractive summarization dataset for multiple low resource languages [15] . The model fine-tunes a pre-trained model mT5 [16] in order to achieve its aim. The model built is openly available for contributions. Even Though multilingual models have been used in NLP tasks previously this is the first model to implement it in the context of abstractive summarization [15]. However due to the computational power the model had to train the model only for five languages from the dataset. They have also stated that due to the less amount of data fed into the model for certain languages the model provides a less evaluation score and the issue of stemming in certain languages is also considered to be a limitation [15]. The research also opens future work on lingual summarization [13].
Timalsina, Paudel and Shahi (2022) [17] proposes an abstractive summarization technique for Nepali text summarization tasks. The authors had to create a dataset from scratch. They have stated that inorder to create the dataset they have scraped data from a Nepali news online website [17]. They propose a method of building a model on a RNN with attention techniques. They have clearly stated that the Nepali language method of implementation is comparatively different to languages like English [17]. The authors claim to be the first to implement an abstractive summarization model for Nepali language [17]. The dataset created is not publicly available but could be gathered upon request. The scarcity of data for a low resource language such as Nepali has been a main limitation of their research. The usage of bi-directional encoders and pointer generation mechanisms [17] are highly recommended for future research by the authors of this research.
Wan and Bansal (2022) [18] propose the idea of presenting summaries that are accurate and factually correct. The authors

were able to use PEGASUS [19] for fine-tuning their datasets and provide improved results during evaluation. They have stated that the sentence errors and token errors generated within the summary have been reduced when compared to SOTA methods [18]. They have also stated that improvement in meaning doesn't always increase the ROGUE scores [18]. They would like to improve the model more so that the ROGUE values also will improve, and it can be used in real world applications [18].

D. Hybrid summarization

Hsu et al. (2018) [6] proposes a research approach of creating a hybrid summarization model with inconsistency loss that includes extractive and abstractive techniques. The model creates a sentence level extractive method and a word level abstractive method for text summarisation [6]. For the model building the authors have used the combination of the best previous models of extractive and abstractive summarization techniques [6]. The authors claim that they have outperformed the SOTA methods with their unified model. To prove the performance of the model the authors have done both automatic and human evaluation. Eventhough though the author has not specifically mentioned future work [6]. This type of an approach havent been used for low - resource languages which makes it a potential area of research. Veenadhari and Bharathi (2022) [20] has proposed a method of hybrid summarization that uses Seq2Seq architecture with RNN. The authors initially planned on implementing it in the legal research domain but due to the scarcity of dataset they had to move onto news summary dataset [20]. They had stated that the technological concept can be used for any domain if the necessary data is fed into the model. The collection of a domain related dataset and the use of pointer generator approaches are stated as future works of the research [20]. Tretyak and Stepanov (2020) [21] research proposes a combined approach of extractive and abstractive techniques for summarizing scientific documents. The authors had used BART [22] for the extractive summarization process and BERT [2] for the abstractive summarization process. The evaluation metrics show that the proposed model outperforms the SOTA methods including the only abstractive model and only extractive model too [21]. As a future work the author would like to try the model on various other domain datasets [21].

E. Technological Review

1) Algorthmic Approaches: This technique was initially and still used only for extractive summarization. The technique involves word frequency and sentence scoring algorithmic approach to extract the most important sentences from a document [23]. TF-IDF and word probability are the most known frequency-based algorithms. LSA is another method that forms a word matrix and compares the cosine similarity between the words and to get the most used sentence and generate the summary [24]. Graph based algorithms are another algorithmic method. This is where the document's words are mapped as a graph to identify the most important words in it [25]. Popular graph methods include PageRank and TextRank algorithms [25].

2) Machine Learning Approaches: When it comes to summarization models were initially built using machine learning. Most explored areas using machine learning are semantic based approaches and query-based summarization [26]. Other than that, some of the researchers used Naive Bayes, Support Vector Machines, K-means clustering and such to build text summarization models. Both extractive and abstractive summarization initially used machine learning techniques until deep learning came into play within the NLP domain and provided more accurate and better results.

3) Deep Learning Approaches: Deep learning approaches have been widely popular in the field of text summarization for a couple of years. Researchers use many different approaches to deep learning in order to achieve the best result for both extractive and abstractive models. Researchers use a lot of pretrained large language models such as T5 [16], BART [22], BERT [2], PEGASUS [27] and transfer learn according to their domain. The use of such models has been very efficient as they have pretrained on large datasets specific for NLP tasks. The Transformers method is the most popular approach that is available for document and text summarization [28].

4) Dataset Preparation: CNN Daily-mail, Xsum, DUC, Gigaword are some of the well collected datasets for summarization tasks. Most of the large language models have been trained on these datasets. But all of these datasets are primarily focussed on the English language. When it comes to low resource datasets, previous researchers mostly create their own datasets by scraping them from websites or manually creating them [17]. Google's translate API has been highly improving by time when it comes to translation tasks. Researchers even state that in many cases the translation includes good grammatical correction. Using real data is the most appropriate method compared to translation but previous researchers who have worked on low resource languages have stated various other ways of gathering data [11]. In Fact, they prove that if the model could score well for such a dataset it will work very well on a well-made dataset [17].

F. Why Document Summarization for Sinhala Language

As mentioned above there is almost 55with an average of 1.5 million users daily [3]. Most of the users consume data in Sinhala language as it is their primary language. This shows that there is a possible number of consumers for applications related to Sinhala language. The survey result gathered in the software requirement specification chapter also shows that there are a lot of Sri Lankans who read Sinhala related documents. As the digital area starts to grow, the information stored starts to increase and the level of consumption reduces [15]. This has already begun with high resource languages such as English and French [11]. This is also predicted to be happening to low resource languages soon. As Sinhala is a low resource language [29] it could be said that due to the increase of data around the internet Sri Lankan people will also stop consuming long Sinhala documents and will be expecting to have shorter simplified versions of them [15]. Models are being implemented for summarizing documents for low resource languages but limitations such as character lengths, scarcity of data, grammatical inconsistencies reduce their level of accuracy but by time they will be improving [8]. This research will be mainly focused on Sinhala document summarization, and it will be the first attempt of doing so. The author identifies the importance of creating such an application as the author foresees that this could be one of the future research areas that would be addressed within the Sinhala research space and provides a great contribution within the Sri Lankan and Sinhala community.

IV. Implementation

A. Proposed Architecture

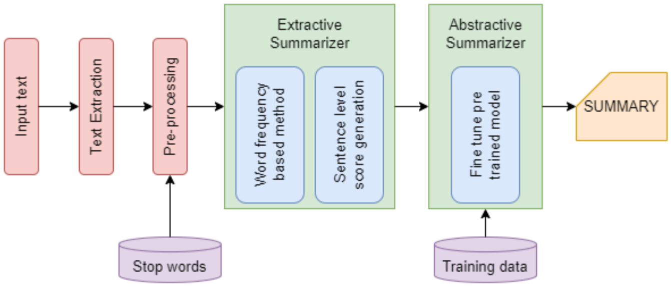

Fig. 1. Proposed Architecture Diagram (Self-composed)

Building a combined approach for summarization which involves both extractive and abstractive techniques has been the goal of this research project. For this the author had to initially identify a valid dataset and prepare it for model training. Once the dataset has been identified the author went on building the extractive model which uses word frequency and sentence scoring algorithms to extract the most important sentences of a document. Once the extractive model extracts the most important sentences it is then sent into the abstractive model for paraphrasing and restructuring it as a human generated summary. This complete process is called the combined approach in which both summarization techniques are used for summarization. Through the implementation the author was able to get the best of both models by extracting keywords (extractive) and restructuring it with own words (abstractive) to generate almost human-like summaries.

B. Extractive model

The extractive model simply uses a traditional approach for extracting the most important sentences of a document. For this five main steps are considered. Tokenization, Word frequency calculation, Sentence scoring,Threshold calculation and Summary generation are followed respectively. Due to this type of aamodel implementation it can be also called as a **frequency based extractive summarization method**.

C. Dataset Preparation

Inorder to train the abstractive model a dataset was created. A subset of the CNN daily mail dataset was translated to create Sinhala articles and summaries. The CNN daily mail dataset includes an approximate of 240000, 14000 and 12000 rows for train, validation, and test respectively. A subset was created where 6000, 2000 and 2000 rows were gathered from the train, validation and test datasets and then translated. For translation of the dataset Google translate API was used. It is to be noted that google translates a maximum of 5000 characters at a time. So, to solve this problem the author skipped rows that had rows more than 4500 characters. The created dataset has been added to the hugging face library. This helps future developers use it for their research and make contributions Hamza-Ziyard/CNN-Daily-Mail-Sinhala.

D. Dataset Preprocessing

Before using the dataset to train the model the dataset was pre-processed. Removal of unwanted columns,rows that has null values,HTML tags,the word CNN in both English and Sinhala, URL links, two consecutive duplicate words [28].

E. Abstractive model

A fine-tuned version of mT5 was used for the training of the abstractive model. The training arguments from Hasan et. al (2021) [15] who was the only one who has attempted in creating an abstractive model for low resource language were considered for the implementation of the model. So the author trained a model with the SOTA training parameters and a model with automated training parameters using Optuna [30]. The batch size was reduced to a maximum of eight for both the models as the author had less resources. Five trials were conducted for generating the best parameters using Optuna due to the lack of resources of the author. The recommended trials is 100 [30]. The results are as follows

TABLE I
SOTA and Optuna training arguments comparision

Arguments	**SOTA arguments**	**Optuna best parameters**
Batch size	8	**4**
Total number of epochs	10	**7**
Learning rate	5.6e-4	5.6e-4
ROUGE-1	20.8581	**21.4151**
ROUGE-2	8.1054	**8.5692**
ROUGE-L	19.8079	**20.1332**

Once the finetuned model has been implmented it was then used for the combined summarizer. The abstractive model has been added to the hugging face library. This helps future developers use it for their research and make contributions. Hamza-Ziyard/sinMT5-tuned

F. Combined model

Finally, the combined approach was built where the input is initially passed into the extractive model which grabs the most important sentences using the frequency based algorithm and then passed into the abstractive model sinMT5-tuned. This combined model then generates a human like meaningful summary of the input document's content.

V. Evaluation

A. Automatic Evaluation

For automatic evaluation ROUGE testing was followed. The ROUGE-N and ROUGE-L equations are as follows.

$$\text{ROUGE-N} = \frac{\text{Number of N-gram Matches}}{\text{Number of N-grams in Reference Summary}}$$

$$\text{ROUGE-L} = \frac{\text{Longest Common Subsequence (LCS) Length}}{\text{Number of Words in Reference Summary}}$$

The test data is from a previous abstractive model that has been tested using the XLSum dataset [15]. The model used a BBC summary dataset for testing. The same summary test data was used for benchmarking The results of the F1 scores are stated in the table below.

TABLE II
BBC News Summary dataset benchmark results

	multilingual XLSUM	**Ours Abstractive**	**Ours Hybrid**
ROUGE-1	27.290	58.7512	**60.2513**
ROUGE-2	13.3815	26.3214	**27.3521**
ROUGE-L	23.4699	**35.5231**	34.2135

It is important to note that the SOTA model used for benchmarking hasn't stated their method of testing. They have just provided the score stating that it is from their test data. Due to time constraints the author picked ten random summaries from the test dataset and generated the above-mentioned ROUGE scores. Due to this the results provided above may not be fully accurate and reliable.

This the reason for extending the testing process for human evaluation to get better reliable results.

B. Human Evaluation

Even though automatic evaluation methods are used to test the performance of a summarization model most of the previous researchers state that human evaluation is the best when it comes to low resource language summarization tasks testing [31]. It is also important to know that automatic evaluation includes a lot of limitations when it comes to low resource language summarization [17]. This reason alone makes human evaluation a high priority.

For conducting the evaluation, the author selected technical experts, domain experts and focus group users. Interviews were done for all the candidates. The number of candidates is stated below.

- Technical Experts - 4
- Domain Experts - 3
- Focus Group - 6

The highlights of their reviews are stated in the following table.

TABLE III
Expert Evaluation results

Expert	**Review**
Technical	NLP for the Sinhala language has been growing for the last couple years which makes it a good research area to work on with. The ability to get the right hyper parameters tuned values for summarization makes the process of summarization much easier and faster than manually tuning to get the best parameters from scratch. The implementation of a hybrid system makes the application better because the output will be a result of the best of both models. The standard of the model is still not as good as English but does solve the aim appropriately. But improving the model with a better accurate and well-crafted dataset can improve the results massively. Training the model on various another domain dataset was highly recommended too. The application speed of generating a summary needs to be improved.
Domain	The results generated from the application are almost grammatically correct. Only in certain cases does the model feed old data to the user. But having an extractive approach was a good way of patching such issues. Improving the time taken for summary generation was also stated
Focus group	Summary generation was meaningful in most cases. Having the opinion of choosing the type of summary to be generated improves the reliability of the application. Summary results perform well for news data. Making it perform well on other domain documents was suggested as an improvement. The model being up to date with real time data and improving time speed for summary generation was also suggested.

VI. Contribution to the Body of Knowledge

A. Problem Domain

The research introduces a Sinhala summarization model that uses a combined approach of both extractive and abstractive approaches for summarization. This helps the Sri Lankan community generate summaries for long Sinhala documents in the way they want which immensely contributes to the problem domain

B. Research Domain

The research shows that by automating the generation of hyper parameter tuning provides better results when compared to using SOTA training arguments [15]. Additionally, the author was able to build a combined approach of summarization that hasn't been built yet for the Sinhala domain. The author has made a translated sub dataset from the openly available CNN Daily Mail dataset which somewhat addresses the scarcity of data [15]. The authors' dataset and abstractive model both are deployed in Huggingface. This makes it open for future researchers to make use of both of and make contributions.

VII. Limitaions and future enhancements

The following are some of the limitations of this research

- The number of trials for hyper parameter tuning to study the best parameters for the model was limited to five as there were computational and time restraints.

- The extractive model had to be a basic algorithmic approach since the dataset used to train isn't fully reliable in generating the best results for extractive summarization.
- Scarcity of real dataset for a Sinhala summarization task made automatic testing and evaluation results not fully accurate.
- ROUGE testing for low resource languages including Sinhala is not accurate during training [15]. This makes evaluation ROUGE scores inaccurate.
- The models have been trained for the news article domain. So, the model might generate some inaccurate results for other domain documents.

The following are some the potential future enhancemenst that could benefit both the research and problem domain areas.

- The abstractive model can be used to expand on other NLP tasks like text classification, generation [22] within the Sinhala language domain.
- Introduction of a question answering system [27]) within this model so that the user could ask questions to explore other content within the document that havent been generated as a summary.
- Improve the model with a real dataset that contains Sinhala articles and Sinhala summary created by Sinhala domain experts.
- Improve model for other document summarization problem domains for Sinhala language. This model is primarily focused on news articles the author would like to extend to other domains like research articles, legal documents and many more.

VIII. Conclusion

This research aimed to implement an application that can be used for summarizaing long Sinhala document content using a combined approach. The research successfully achieved its aim by implementing a combined model with both extractive and abstractive models. It also proves that the model provides better results when compared to previous work. Both automatic and human evlaution results show that the model achieve industry levels to provide an efficient application. Usage of the model to contribute for other NLP research areas could be a great future enhancement. While there is still room for improvement for now it can be concluded that it meets necessary requirements. All implemented models and code is made publicly available for future contributions to both the Sinhala and NLP domains. Overall, this research can be concluded by stating that it masively contributes to the field of Sinhala document summarization.

References

[1] A. Geitgey, "Natural Language Processing is Fun!" Sep. 2020. [Online]. Available: https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[3] "Digital 2022: Sri Lanka," Feb. 2022. [Online]. Available: https://datareportal.com/reports/digital-2022-sri-lanka

[4] L. Gonçalves, "Automatic Text Summarization with Machine Learning — An overview," Sep. 2021. [Online]. Available: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25

[5] G. Carenini and J. C. K. Cheung, "Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality," in *Proceedings of the Fifth International Natural Language Generation Conference*. Salt Fork, Ohio, USA: Association for Computational Linguistics, Jun. 2008, pp. 33–41. [Online]. Available: https://aclanthology.org/W08-1106

[6] W.-T. Hsu, C.-K. Lin, M.-Y. Lee, K. Min, J. Tang, and M. Sun, "A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 132–141. [Online]. Available: https://aclanthology.org/P18-1013

[7] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, "MSˆ2: Multi-Document Summarization of Medical Studies," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 7494–7513. [Online]. Available: https://aclanthology.org/2021.emnlp-main.594

[8] V. Dhananjaya, P. Demotte, S. Ranathunga, and S. Jayasena, "BERTifying Sinhala – A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification." arXiv, 2022, version Number: 2. [Online]. Available: https://arxiv.org/abs/2208.07864

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 2019, arXiv:1910.13461 [cs, stat] version: 1. [Online]. Available: http://arxiv.org/abs/1910.13461

[10] K. A. M. P. Rathnasena, K. M. S. J. Kumarasinghe, D. T. P. Paranavitharana, D. V. A. U. Dayarathne, and L. Ranathunga, "Summarization based approach for Old Sinhala Text Archival Search and Preservation," in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Sep. 2018, pp. 182–188, iSSN: 2472-7598.

[11] P. Deshpande and S. Jahirabadkar, "Study of Low Resource Language Document Extractive Summarization using Lexical chain and Bidirectional Encoder Representations from Transformers (BERT)," in *2021 International Conference on Computational Performance Evaluation (ComPE)*, Dec. 2021, pp. 457–461.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[13] J. Zhu, Q. Wang, Y. Wang, Y. Zhou, J. Zhang, S. Wang, and C. Zong, "NCLS: Neural Cross-Lingual Summarization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3054–3064. [Online]. Available: https://aclanthology.org/D19-1302

[14] A. Sakhadeo and N. Srivastava, "Effective extractive summarization using frequency-filtered entity relationship graphs," Oct. 2018, arXiv:1810.10419 [cs]. [Online]. Available: http://arxiv.org/abs/1810.10419

[15] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar, "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4693–4703. [Online]. Available: https://aclanthology.org/2021.findings-acl.413

[16] P. Mishra, "Understanding T5 Model : Text to Text Transfer Transformer Model," May 2021. [Online]. Available: https://towardsdatascience.com/understanding-t5-model-text-to-text-transfer-transformer-model-69ce4c165023

[17] B. Timalsina, N. Paudel, and T. B. Shahi, "Attention based Recurrent Neural Network for Nepali Text Summarization," *Journal of Institute of*

Science and Technology, vol. 27, no. 1, pp. 141–148, Jun. 2022. [Online]. Available: https://www.nepjol.info/index.php/JIST/article/view/46709

[18] D. Wan and M. Bansal, "FactPEGASUS: Factuality-Aware Pre-training and Fine-tuning for Abstractive Summarization," 2022. [Online]. Available: https://www.semanticscholar.org/reader/5f50a876e1f323598df423475a9cb4c01bd4b44f

[19] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," Jul. 2020, arXiv:1912.08777 [cs]. [Online]. Available: http://arxiv.org/abs/1912.08777

[20] G. V. Veenadhari and B. R. Bharathi, "IMPROVING HYBRID SUMMARIZATION BY USING ABSTRACT AND EXTRACT MODEL," vol. 10, no. 9, 2022.

[21] V. Tretyak and D. Stepanov, "Combination of abstractive and extractive approaches for summarization of long scientific texts," Jun. 2020, arXiv:2006.05354 [cs]. [Online]. Available: http://arxiv.org/abs/2006.05354

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 2019, arXiv:1910.13461 [cs, stat] version: 1. [Online]. Available: http://arxiv.org/abs/1910.13461

[23] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text Summarization Techniques: A Brief Survey," Jul. 2017, arXiv:1707.02268 [cs]. [Online]. Available: http://arxiv.org/abs/1707.02268

[24] J. Steinberger and K. Ježek, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation."

[25] R. Mihalcea, "Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 170–173. [Online]. Available: https://aclanthology.org/P04-3020

[26] Rahul, S. Adhikari, and Monika, "NLP based Machine Learning Approaches for Text Summarization," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Mar. 2020, pp. 535–538.

[27] J. Zhu, Q. Wang, Y. Wang, Y. Zhou, J. Zhang, S. Wang, and C. Zong, "NCLS: Neural Cross-Lingual Summarization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3054–3064. [Online]. Available: https://aclanthology.org/D19-1302

[28] K. Divya, K. Sneha, B. Sowmya, and G. S. Rao, "Text Summarization using Deep Learning," vol. 07, no. 05, 2020.

[29] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6098–6111. [Online]. Available: https://aclanthology.org/D19-1632

[30] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," Jul. 2019, arXiv:1907.10902 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1907.10902

[31] C.-H. Lee, A. Siddhant, V. Ratnakar, and M. Johnson, "DOCmT5: Document-Level Pretraining of Multilingual Language Models," *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 425–437, 2022, conference Name: Findings of the Association for Computational Linguistics: NAACL 2022 Place: Seattle, United States Publisher: Association for Computational Linguistics. [Online]. Available: https://aclanthology.org/2022.findings-naacl.32