Multilingual Transformer-Based Summarization for Low-Resource Nepali News Articles

Bidhya Bhattarai[1,*], Biplav Chaudhary[2], Punam Bashyal[3], Sajita Khadka[4], Krishnanand Badu[5]
[1,2,3,4,5]Himalaya College of Engineering, Tribhuvan University (TU), Lalitpur, Nepal
*Corresponding author: Bidhya Bhattarai, hce078bct010@hcoe.edu.np

Abstract

The study refines an abstract Nepali news summarization system using natural language processing (NLP). The powerful multilingual T5 (mT5) model was fine-tuned in the collected data set. Pre-processing steps, including tokenization, punctuation removal, and special character removal, were applied to enhance performance. Using supervised learning, the model was trained to reduce overfitting. Evaluation was conducted using the ROUGE metric to assess the quality of the generated summaries. The extensive text is then provided to users in the form of concise and meaningful summaries, preserving the core meaning of the original content. The news articles were also extracted using an API, and the summaries are displayed accordingly. This paper highlights the transformer- based model for low-resource languages like Nepali. Moving forward, the plan is to secure more powerful computational resources and improve the scalability of the generated summaries.

Keywords: Abstractive – Low-Resource Language – mT5 – NLP – Transformer

1. Introduction

The advancements in automatic text summarization tools have primarily focused on high resource languages such as English, which of ten utilize extractive approaches. However, these methods are not well suited for low resource languages like Nepali due to the limited availability of large-scale pretrained models. As a result, Nepali has not gained significant attention in the field of Natural Language Processing (NLP). In the information-driven world, it is becoming difficult for readers to quickly identify and consume key information from news articles. This presents a clear need for effective summarization tools for the Nepali language. To address these challenges, this project aims to develop an abstractive Nepali News Summarization System that generates concise and human like summaries that preserve the important meaning of the original content. Far from extractive summarization, which just selects sentences from the input text, our approach focuses on generating important sentences that capture the essence of the input. Such a system not only helps readers save their time but also enhances access to important information within limited internet bandwidth. The primary objective is to design a summarization system that takes full length Nepali news articles as input and produces accurate, coherent, readable, and meaningful summaries. This tool has potential applications across multiple domains, including journalism, education, and media monitoring, and offers an efficient means of digesting and retrieving essential news insights

2. Literature Review

Automatic text summarization is an essential task within Natural Language Processing (NLP), designed to produce concise summaries that emphasize key information from lengthy texts and documents. Although significant advances have been made for high-resource languages such as English, this field remains inadequately explored for low-resource languages such as Nepali. This shortfall can be attributed to different challenges, including morphological complexity and the scarcity of available datasets [10].

Initial efforts in text summarization largely utilized extractive methods, which involve selecting highlighted words and segments directly from the original text. Among various approaches, TextRank, a graph-based ranking model, has gained popularity due to its effectiveness in summarizing. In this model, sentences are depicted as vertices (nodes) in a graph, with nodes assigned weights based on their similarity to each other [3]. Frequency-based methods have also been prevalent in identifying terms that appear most frequently (Lin, 2004). However, these approaches frequently prove insufficient for low-resource languages such as Nepali due to their intricate morphology and rich vocabulary [6].

To improve summarization in low-resource languages, [3] developed a combined method that integrates TextRank with topic modeling to summarize Nepali documents. Their results demonstrated the feasibility of using Text Rank for Nepali. However, extractive methods often fail to generate concise and semantically meaningful summaries. A notable study by [8] introduced an attention-based RNN model with LSTM units for abstractive summarization of Nepali news articles. Although the model showed potential, its performance was modest due to data limitations and the intrinsic complexity of the language.

The introduction of the Transformer architecture revolutionized NLP through the use of self-attention mechanisms [9]. This innovation led to the development of powerful models such as BERT [1] and T5 [7], which significantly outperform previous approaches in language understanding and text generation tasks. Although most transformer-based research has focused on high-resource languages, the availability of multilingual models like mT5 has opened new opportunities

Article History:
Submission: 2025-05-30, Revision: 2025-09-21, Acceptance: 2025-11-17

for low-resource languages such as Nepali [4].

Alongside model advances, evaluation metrics have played a crucial role in benchmarking summarization quality. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures overlap between system-generated and reference summaries, and is widely used for summarization tasks [5]. BLEU (Bilingual Evaluation Understudy) evaluates fluency and accuracy based on n-gram matches (Papineni et al., 2002). These metrics enable fair comparisons between extractive, neural, and transformer-based approaches.

Recent studies have also emphasized the role of multilingual and cross-lingual datasets. The XL-Sum dataset [6] provides multilingual news summaries, while WikiLingua [4] offers parallel cross-lingual summaries from instructional content. Such datasets have opened doors for low-resource languages like Nepali to leverage multilingual pre-trained models. [10] further demonstrated that multilingual transformers can adapt across languages for cross lingual summarization, highlighting their potential for resource-scarce languages.

This study advances Nepali text summarization by leveraging the multilingual mT5 model, achieving notable improvements over traditional extractive methods and early neural architectures. Unlike extractive techniques such as TextRank, which select existing sentences from the input, the mT5 model generates new sentences that better preserve semantic meaning and contextual flow. In contrast to attention-based RNN models, which often struggle with long-range dependencies and vocabulary limitations, the transformer-based mT5 architecture produces more coherent and fluent summaries [7].

Fine-tuning was performed using a large-scale Nepali news dataset obtained from [2], which contains over 17,000 articles sourced from platforms such as Setopati. Preprocessing steps, including SentencePiece tokenization, stopword removal, and normalization, were applied to improve data quality and training effectiveness. These enhancements address the key challenges of Nepali morphological complexity and limited NLP resources, demonstrating the potential of the model in low resource language settings.

Overall, the mT5 model outperformed traditional extractive methods (such as TextRank) and RNN-based approaches in generating more coherent and semantically meaningful summaries. While there is still room for improvement—particularly in capturing deeper context and complex word relationships—the current results demonstrate the model's potential and provide a strong foundation for future enhancement. To facilitate practical use, a web-based application was also developed with features including user login, article submission, summary display, feedback collection, and word count tracking. The system supports both automatically extracted news content via API and user submitted text, providing a versatile platform for Nepali language news summarization

2.1 Research Objectives

- To develop an abstractive summarization model for Nepali news articles using the mT5 transformer.
- To pre process and fine-tune the model using a large-scale Nepali news dataset.
- To deploy a user-friendly web-based application for real-time Nepali news summarization.

3. Methodology

The diagram shows the architecture of the mT5 (Multilingual T5) model, a transformer based sequence to sequence model. It starts with to kenizing input text using Sentence Piece, followed by embedding tokens with positional information. The encoder processes the input using self attention or multi head attention and feed-forward layers, while the decoder generates output using masked self attention and similar feed-forward blocks. Atlast, the output layer applies a linear transformation and softmax to produce predictions.

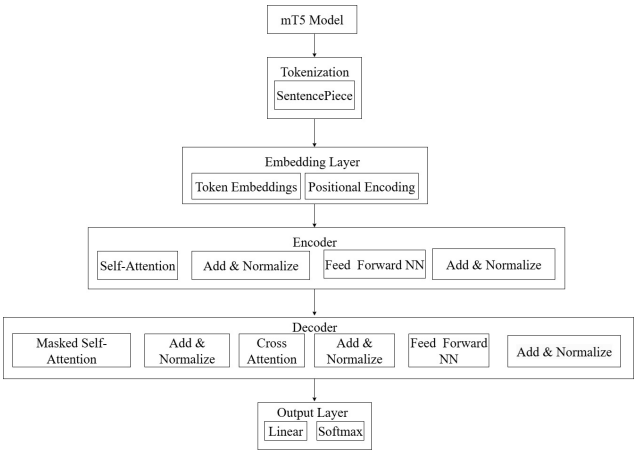

Figure 1. mT5 Model Architecture

3.1 Dataset and Preprocessing

The dataset contains news articles in Nepali along with their corresponding summaries. It covers a broad range of topics, including national news, economy, sports, and events occurring in Nepal. This dataset consists of 17,312 rows with two columns: one named 'article' and the other containing the respective summary, called 'article summary'. The data has been split into three parts: a training set with 4,038 rows, a validation set with 505 rows, and a test set with 505 rows. This split enables efficient model training and evaluation, ensuring that the model can generalize well to unseen data while being fine-tuned on the validation set.

Figure 2. Dataset Used

Data pre processing is a crucial step in preparing text data for training the mT5 model, ensuring it is compatible and optimized for effective learning. This process involves several steps starting with cleaning the data to remove irrelevant or special characters. Next, tokenization breaks down the text into smaller units such as words or subwords, making it suitable for a transformer-based model. Word2Vec is then applied to generate dense vector representations for each word, capturing semantic meanings and relationships between words.

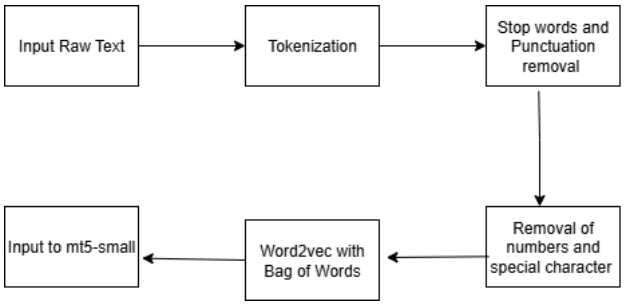

Figure 3. Block diagram of Preprocessing

3.2 Input Encoding

The input text is initially tokenized using methods such as Sentence Piece, which divide the text into smaller units like words or subword tokens. Each token is then mapped to a unique numerical representation known as an embedding. These embeddings capture the meaning and contextual relationships of the input tokens.

3.3 Encoder

The encoded input tokens are passed through a stack of identical encoder layers. Each encoder layer consists of two sub layers: a self attention mechanism and a feed forward neural network. The self attention mechanism allows the model to identify dependencies between different tokens in the input sequence. Attention weights are computed to highlight the relative importance of each token in the context of others. These attention weights guide the model to focus on the most relevant parts of the input, enhancing its ability to capture detailed patterns and relationships within the sequence.

3.4 Decoder

The encoder layers process the input tokens concurrently, allowing the model to capture contextual dependencies across the entire sequence. Each decoder layer consists of three sub layers: a masked self-attention mechanism, an encoder-decoder attention mechanism, and a feed-forward neural network. The masked self-attention mechanism enables the decoder to attend only to earlier tokens in the target sequence during generation. This prevents the model from accessing future tokens during inference and ensures attention to relevant prior information when predicting the next token. The encoder-decoder attention mechanism enables the decoder to focus on important parts of the encoder's output, generating con- textually appropriate and semantically accurate output tokens based on the learned representations.

3.5 Outer Layer

The output of the decoder is first passed through a linear transformation, projecting it into a higher-dimensional space to capture complex relationships and patterns. This transformed output is then processed using a softmax activation function, which converts it into a probability distribution over the vocabulary. Each value in this distribution represents the likelihood of a particular token being the next in the summary sequence. During training, the next token is typically sampled based on these predicted probabilities, enabling the generation of coherent and contextually appropriate summaries.

3.6 Hyperparameters

The following hyper parameters were used to train and evaluate the mT5 model:

- **Epochs:** 30 – The model was trained over the dataset for 30 complete passes.
- **Batch size:** 1 – Weights were updated after every single data point.
- **Learning rate:** 1 × 105 – A small step size for weight updates.

3.7 Evaluation

3.7.1 BLEU

BLEU (Bilingual Evaluation Understudy) score is a widely used metric for machine translation tasks, where the goal is to automatically translate text from one language to another. It was proposed as a way to assess the quality of machine generated translations by comparing them to a set of reference translations provided by human translators. BLEU score calculates the precision of ngrams in the machine generated translation by comparing them to the reference translations. The precision is then modified by a brevity penalty to account for translations that are shorter than the reference translations. The formula of BLEU

$$BLUE = BP * exp(\sum(pn)) \tag{1}$$

3.7.2 ROUGE

ROUGE(Recall-Oriented Understudy for Gisting Evaluation) is a widely used evaluation metric in natural language processing (NLP) for assessing the quality of automatically generated summaries. It measures the overlap between the generated summary and a reference (gold standard) summary, which is typically created by human annotators. The overlap is computed based on ngrams (word sequences), with common values being unigrams (single words) and bigrams (word pairs). The ROUGE score ranges from 0 to 1, with a score of 1 indicating a perfect match.The formula for the ROUGE score between an automatic summary S and reference summary R can be expressed as:

$$\sum N_n = \frac{\text{count}(n\text{-gram}(S) \cap n\text{-gram}(R))}{\text{count}(n\text{-gram}(R))} \tag{2}$$

$$\text{Precision} = \frac{\text{number of } n\text{-grams found in candidate and ref}}{\text{number of } n\text{-grams found in ref}} \quad (3)$$

$$\text{Recall} = \frac{\text{number of } n\text{-grams found in candidate and ref}}{\text{number of } n\text{-grams found in ref}} \quad (4)$$

4. Results and Discussion

4.1 Rouge value and BLEU score for mT5 Model

The Table 1 presents the evaluation metrics for the summarization model. ROUGE scores measure the overlap of n-grams between the generated and reference summaries, while the BLEU score evaluates the accuracy based on precision. ROUGE-1(0.4176) measures the overlap of unigrams between the generated summary and the reference summary.

Table 1. Rouge Value and BLEU Score

Metric	Score
ROUGE-1	0.4176
ROUGE-2	0.1867
ROUGE-L	0.4176
BLEU SCORE	0.3282

A score of 0.4176 indicates that approximately 41.76% of the words in the generated summary match those in the reference summary.ROUGE- 2(0.1867) measures the overlap of bigrams(pairsofconsecutivewords)between the generated and reference summaries. A score of 0.1867 indicates that about 18.67% of the bigrams match. ROUGE-L(0.4176) evaluate show well the model captures the overall struc- ture and flow of the text. A score of 0.4176

indicates that the model performs similarly in capturing the overall structure as it does with unigrams, suggesting consistency in generat- ing coherent summaries. The BLEU SCORE (0.3282) score of 0.3282 indicates that ap- proximately 32.82% of the generated summary matches the reference in terms of word choice and phrasing. This score reflects the model's ability to produce linguistically accurate and relevant summaries.

4.2 Loss diagram and Accuracy comparison

The accuracy comparison graph illustrates the training and validation accuracy of the mT5 model over 30 epochs. Training accuracy im- proves steadily with time, indicating that the model is learning effectively. The validation accuracy also increases initially but shows slower progress compared to the training accu- racy, with fluctuations observed after a certain number of epochs. The loss comparison graph illustrates the training and validation loss of the mT5 model over 30 epochs. The training loss decreases significantly over the epochs, while the validation loss decreases initially but begins to increase after 12 epochs. The training loss continues to decline steadily, indicating ongoing learning, while the validation loss stabilizes and then increases, suggesting a divergence in performance between the training and validation datasets.

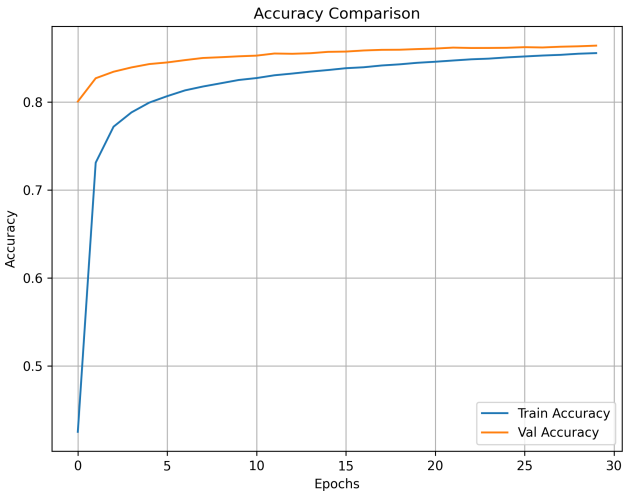

Figure 4. Accuray Comparison

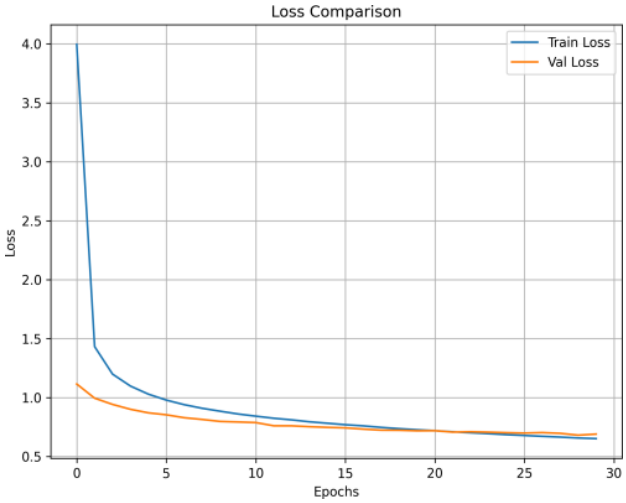

Figure 5. Loss Diagram for MT5 Model

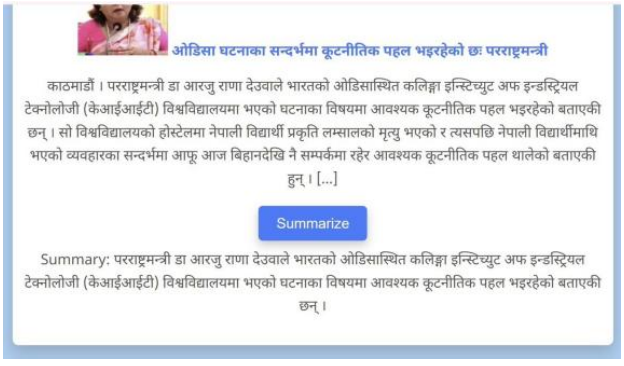

Figure 6. News Summarization of Nepali News Website

5. Conclusion

The project Nepali News Summarization aimed to generate concise yet comprehensive summaries accurately reflecting

Figure 7. Any Nepali Summarization

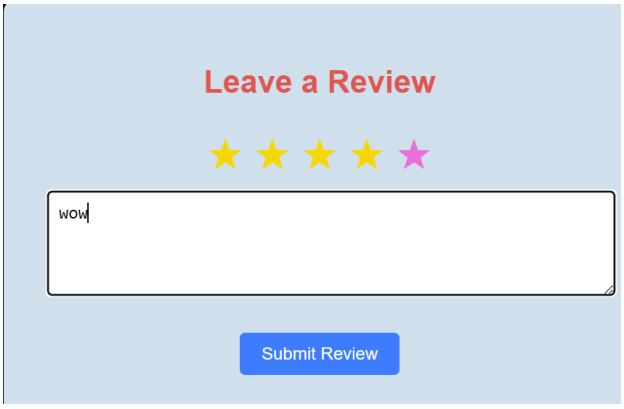

Figure 8. Review of Summary

the core content of the news article using MT5 model which is the powerful multilingual transformer based model. Fine tuned specially for Nepali text summarization utilized a publicly available dataset from Hugging Face,applying preprocessing techniques to enhance model performance. The model's performance was tested using evaluation metrics (ROUGE and BLEUscore) that demonstrated how the well formed and accurate the sentences are compared to a reference summary. The system is designed to summarize both extracted news articles and user provided text, making it a versatile tool for information consumption even for low-resource languages.

Future Enhancement

Future research should build diverse datasets covering various Nepali domains and dialects, develop Nepali specific models, and blend extractive and abstractive methods. Adding features like English translation, a Python module for text extraction, a browser extension for easy summarization, or multimodal summarization (mixing text with speech or images) could make these tools more practical and widely used in Nepal's growing digital landscape. Currently, the lack of Nepali specific models poses a challenge, but we remain hopeful for future developments in this area. Additionally, we faced difficulties in gathering a large and diverse dataset of Nepali articles and their summaries, which is crucial for enhancing model performance. Expanding our dataset is a top priority, as it would improve the training process and the quality of the summaries generated.

Acknowledgment

We take this opportunity to express our heartfelt gratitude to everyone who supported and guided us throughout the development of our minor project, Nepali News summarization. First and foremost, we sincerely extend our hearfelt gratitude to our project supervisor, Er. Krishnanand Badu, for his invaluable guidance, constant encouragement, and constructive feedback, which were instrumental in the successful completion of the project. We would like to convey our sincere and heartfelt thanks to the minor project coordinator, Er.Shiva Raj Luitel, and Er.Narayan Adhikari Chhetri, for providing us with essential resources and constant encouragement, which equipped us with the knowledge and skills required to complete this project. We would also like to express our gratitude to Himalaya College of Engineering for providing the necessary resources and a conducive learning environment. Finally, we acknowledge the importance of collaborative efforts, dedication, and hard work that contributed to the successful realization of our project goals

Conflict of Interest

The authors declare no conflict of interest.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[2] Hugging Face, "Someman/news_nepali dataset," *Hugging Face Datasets*, 2023. [Online] Available:

[3] S. Kumar, N. Singh, S. Agarwal, and S. Bhattacharyya, "Extractive text summarization using TextRank and topic modeling for low-resource languages," *IEEE Access*, vol. 8, pp. 109934–109944, 2020, doi: 10.1109/ACCESS.2020.3002192.

[4] F. Ladhak, E. Durmus, C. Cardie, and K. McKeown, "WikiLingua: A new benchmark dataset for cross-lingual and multilingual abstractive summarization," *arXiv preprint arXiv:2010.03093*, 2020. [Online]. Available: https://arxiv.org/abs/2010.03093

[5] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL-04 Workshop Text Summarization Branches Out*, 2004, pp. 74–81. Available: https://aclanthology.org/W04-1013

[6] S. Narayan, S. Reddy, R. Tang, and R. McDonald, "The XL-Sum dataset for multilingual abstractive summarization," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, 2022, pp. 3009–3026, doi: 10.18653/v1/2022.acl-long.216.

[7] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020. Available: http://jmlr.org/papers/v21/20-074.html

[8] B. Timalsina, N. Paudel, and T. B. Shahi, "Attention-based recurrent neural network for Nepali text summarization," *J. Inst. Sci. Technol.*, vol. 27, no. 1, pp. 11–20, 2022, doi: 10.3126/jist.v27i1.46709.

[9] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008. Available: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[10] G. I. Winata *et al.*, "Learning fast adaptation on cross-lingual sum-

marization across low-resource languages," in *Proc. 2021 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2021, pp. 758–766, doi: 10.18653/v1/2021.naacl-main.62.