

Text Summarization for Indian Languages: Finetuned Transformer Model Application ^{*}

V Ilanchezhiyan¹, R Darshan², E M Milin Dhithshithaa³ and B Bharathi⁴

^{1 2 3 4} Department of CSE, Sri Siva Subramaniya Nadar College of Engineering, Rajiv Gandhi Salai, Chennai, Tamil Nadu, India

Abstract

Summarization, like distilling ancient wisdom, has stood the test of time. From ancient storytelling to modern contexts like meetings, the act of condensing information remains relevant. This abstracts itself exemplifies the ongoing use of summarization. In today's digital age, where technology mimics once-exclusive human tasks, text summarization is a notable example. Natural Language Processing (NLP) and AI models have automated this skill, though attention to Indian languages remains sparse. This paper explores the work on ILSUM (Indian Language Summarization) in the FIRE 2023 task, comparing existing models. Our m2023 model achieved the second position for English in FIRE 2023 rankings. We used mT5-small and mT5-base with fine tuned T5-base: m2023-t5-base, excelling in generating precise summaries.

Keywords

Generated Summary, Indian Languages, Pre-Trained Model, m2023-t5-base, Hindi, Gujarati, Bengali

1. Introduction

Natural Language Processing (NLP), a field at the intersection of computer science and artificial intelligence (AI), integrates computational linguistics with statistical machine learning and deep learning models. It focuses on developing rule-based models for human language, enabling computers to process both voice data and text. This capability allows computers to read text, comprehend speech, and derive meaning from it. NLP breaks down language into tokens, seeking to understand the relationships between these tokens. The spectrum of NLP tasks includes sentiment analysis, word sense disambiguation, grammatical tagging, content categorization, text summarization, topic discovery and modelling, speech-to-text, and vice versa, among others. These tasks encounter challenges in achieving accuracy due to the inherent ambiguities in human language, as well as the diversity of languages, the use of metaphors, sarcasm, idioms, and various grammatical exceptions.

Text summarization stands as a pivotal application within Natural Language Processing (NLP), serving the purpose of condensing large volumes of textual information into concise and meaningful summaries. This process, often referred to as Automatic Text Summarization (ATS), is crucial in extracting the most salient information from extensive documents, enabling


Forum for Information Retrieval Evaluation, December 15-18, 2023, India

[†] These authors contributed equally.

✉ ilanchezhiyan2110023@ssn.edu.in (V. Ilanchezhiyan); darshan2110024@ssn.edu.in (R. Darshan); milin2110126@ssn.edu.in (E. M. M. Dhithshithaa); bharathib@ssn.edu.in (B. Bharathi)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

efficient content digestion and comprehension. As information overload becomes an increasingly prevalent challenge in the digital age, the significance of text summarization has grown, offering a solution to distil key insights from vast datasets. These are several papers for utilising pre-trained models for text summarisation. [1] While numerous NLP tools and approaches address these challenges effectively, their applicability is often constrained when dealing with low-resource languages. Text summarization, a key application of NLP techniques, involves processing extensive digital text from sources like articles, magazines, or social media. The goal is to generate concise summaries and synopses for indexes, research databases, or time-constrained readers. Automatic Text Summarization refers to the computer-driven execution of this task using algorithms and programs. The text summarization algorithms are compared and contrast in this paper [2].

1.1. Summarization Techniques

Various techniques are employed in text summarization, each leveraging advancements in computational linguistics, machine learning, and deep learning. Rule-based models, which rely on predefined linguistic rules, are foundational in summarization. These models extract important sentences or phrases based on syntactic and semantic structures. Statistical methods, such as frequency analysis and TF-IDF (Term Frequency-Inverse Document Frequency), quantify the importance of words and sentences to identify key content. Machine learning techniques, particularly supervised methods, utilize labelled data to train models to discern relevance and generate summaries. Deep learning models, including recurrent neural networks (RNNs) and transformers, have revolutionized summarization by capturing intricate contextual relationships within the text.

1.2. Types of Models

Text summarization encompasses diverse approaches, with extractive and abstractive summarization being the primary paradigms. Extractive summarization involves selecting and rearranging existing sentences from the source text to form a summary. This method relies on sentence importance scores calculated through various algorithms, such as graph-based models or machine learning classifiers. Abstractive summarization, on the other hand, goes beyond extraction by generating new sentences that capture the essence of the source text. This process requires an understanding of the content and often involves complex NLP models, such as transformers, that can paraphrase and rephrase information to create coherent summaries. In recent years, pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have gained prominence in text summarization. These models, trained on vast amounts of diverse text data, demonstrate superior language understanding and generation capabilities. Fine-tuning these models for summarization tasks has yielded state-of-the-art results in producing human-like summaries. In this dynamic landscape, the exploration of novel architectures, hybrid models, and linguistic innovations continues to drive advancements in text summarization. This comprehensive overview aims to elucidate the multifaceted nature of summarization, showcasing its evolution from traditional rule-based systems to cutting-edge deep learning approaches, and highlighting

the ongoing quest for more effective and linguistically nuanced models. The remainder of the paper is organised as follows: Section 2 discusses the literature survey of related works. The descriptions of data and the proposed methods are detailed in Sections 3 and 4 respectively. Section 5 underscores the achieved results and presents a thorough analysis of the performances of each model. Section 6 concludes the paper and contains a discussion about future research.

2. Related Work

The literature on Indian language summarization spans various languages, with a focus on techniques and approaches tailored to specific linguistic nuances. A work [3] employs neural network models for abstractive summarization in Hindi, leveraging the Transformer architecture for improved contextual understanding.

In the context of English, research [4] delves into the application of BERT-based models for summarization tasks, demonstrating their effectiveness across diverse datasets. Another study [5] explores the use of reinforcement learning in abstractive summarization, showcasing advancements in generating coherent and concise summaries.

For Bengali, a noteworthy paper [6] investigates the challenges and opportunities in Bengali language summarization. The work emphasizes the need for language-specific models to capture the intricacies of Bengali, showcasing the development of a summarization system tailored to this language's linguistic characteristics.

In Gujarati, research [7] focuses on leveraging pre-trained language models for summarization tasks. The study provides insights into the adaptability and performance of such models in the context of Gujarati language summarization.

To further enrich the literature survey, a comparative analysis of these studies across Hindi, English, Bengali, and Gujarati reveals the diverse methodologies employed, ranging from graph-based approaches to neural networks and pre-trained models. These works collectively contribute to the ongoing advancements in Indian language summarization, acknowledging the linguistic diversity and unique challenges posed by each language.

The previous papers [8][9][10][11][12] on the shared tasks organised by FIRE ranging from different approaches and models provided various perspectives in visualising a solution to the problem.

In the exploration of the task, dataset, and methodologies employed by various participants, references to the shared task papers offer comprehensive insights. The work by [13] delves into the intricacies of Indian Language Summarization at FIRE 2023, encompassing diverse approaches and dataset considerations. Similarly, the collaborative effort outlined in [14] provides a comprehensive overview of the Second Shared Task on Indian Language Summarization (ILSUM 2023), elucidating the strategies adopted by multiple teams.

3. Experiment Dataset

The following section provides a detailed description of the data used in this study as well as the preprocessing techniques employed. The undertaken task has also been discussed comprehensively.

3.1. Data Description

The dataset for this task comprises approximately 15,000 news articles in each language, drawn from leading newspapers in the country. The objective is to generate a meaningful fixed-length summary, be it extractive or abstractive, for each article. Notably, the dataset introduces a unique challenge of code-mixing and script mixing, where phrases from English are commonly integrated into news articles, even when the primary language is an Indian language.

A distinctive feature of this dataset is its inclusion of examples exhibiting code-mixing in both headlines and articles, reflecting the common practice of borrowing English phrases within Indian language content. For instance:

"IND vs SA, 5मी T20 तसवीरोमां: वरसाटे विषन बनी मज्ज बग़ाडी"
(India vs SA, 5th T20 in pictures: rain spoils the match)

"LIC के IPO में पैसा लगाने वालों का टूटा दिल, आई एक और नुकसानदेह खबर"
(Investors of LIC IPO left broken hearted, yet another bad news).

Figure 1: Illustrating code-mixing in the dataset

The dataset is structured into four CSV files: english-train.csv, hindi-train.csv, gujarati-train.csv, and bengali-train.csv. Each file contains columns for articles and summaries, providing a comprehensive foundation for training and evaluating models on the task of generating meaningful summaries for news articles in diverse Indian languages, while accommodating the intricacies of code-mixing and script mixing.

3.2. Task Description

Generate concise and meaningful fixed-length summaries for news articles in multiple Indian languages, considering the challenge of code-mixing and script mixing. The dataset includes articles and headline pairs from leading newspapers in English, Hindi, Gujarati, and Bengali.

3.3. Data Pre-processing

The dataset employed in this study comprises articles and summary in the languages English and Hindi.

Special characters and punctuations occur frequently in the dataset. No preprocessing step is required to scan through a list of special characters and replace special characters with a space in the text data of the training, validation, and test datasets as the context is important in generating summaries. Further, entries with missing values or labels were not present as the provided dataset was pre-processed before. Data Cleaning: The Given dataset didn't have any null characters, empty lines or newline characters. Although the data was mixed with html or xml entity codes. Those codes had to be replaced with its corresponding characters such as "’" with "'", "–" with "-", "“" with "\"", "”" with "" and

"‘", "'". Without this pre-processing step the model generated random words along with html or xml entity codes.

Before pre-processing: Heavy rains in Chennai’ Pondicherry and Kerala...

After pre-processing: Heavy rains in Chennai, Pondicherry and Kerala

4. Proposed Methodology

This section details in-depth explanations for each of the experiments conducted for the shared task. Figure 1 depicts the general flow of the process.

4.1. Multilingual Summarization with T5-base and Translation

In the initial stages of this project, the selection of the T5-base model as the primary summarization model was informed by a meticulous comparative analysis. This analysis demonstrated that T5-base consistently outperformed its smaller counterpart, T5-small, offering more dependable and contextually accurate summaries. The decision to leverage T5-base aligns with the project's overarching objective of achieving high-quality summarization results across the linguistically diverse landscape of Indian English and Hindi.

4.2. Comparative Analysis for Model Selection

To establish the efficacy of T5-base, a comparative analysis was conducted, considering various summarization models. The results indicated that T5-base consistently demonstrated superior performance, making it the model of choice for this multilingual summarization task. The model's ability to handle diverse linguistic nuances and generate contextually accurate summaries influenced this selection.

4.3. Evaluation Metrics

The effectiveness of the summarization models is assessed using two key metrics: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BERT (Bidirectional Encoder Representations from Transformers) scores. These metrics provide a comprehensive evaluation of the generated summaries:

ROUGE Scores: ROUGE evaluates the overlap between the generated summaries and reference summaries, offering a quantitative measure of the content's quality. It assesses the precision of the model in capturing essential information from the source articles.

BERT Scores: BERT scores gauge the semantic similarity between the generated and reference summaries. This metric delves into the contextual understanding of the model, providing insights into how well the summarization captures the underlying meaning of the source text.

4.4. Main Workflow

The primary workflow of the proposed methodology involves the following steps:

Fig 1. Flow Diagram of the Main Workflow

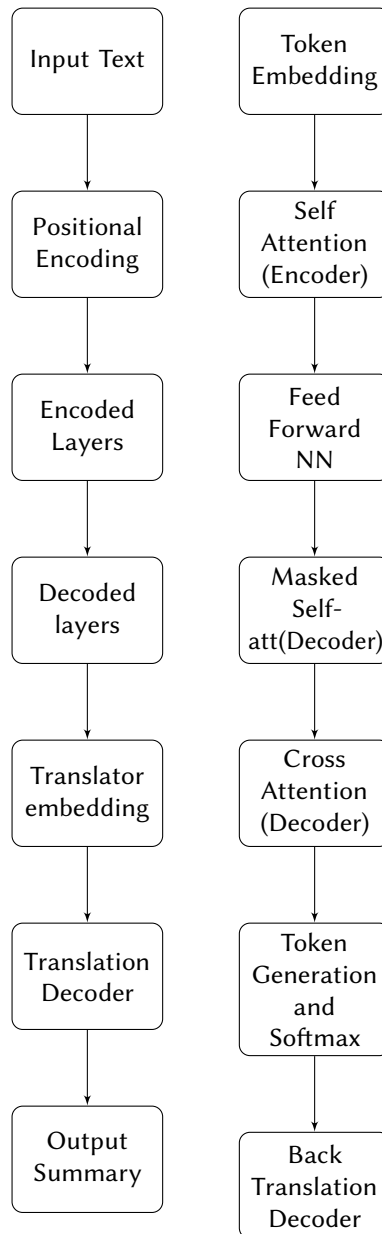


Figure 2: Fig 1. Flow Diagram of the Main Workflow in T5-base

- **Translation to English:** Articles in Indian languages (Hindi) are translated into English using a language translation model. This step aims to leverage the extensive English datasets available for training the summarization model.
- **Summarization with T5-base:** The translated articles are fed into the T5-base model, they are fine-tuned using the FIRE 2023 dataset specific to the summarization task. The model generates English summaries that capture the key information from the translated

articles and generate summaries in English that need to be back translated.

- **Back-Translation:** The generated English summaries undergo a subsequent step where they are translated back into the original Indian languages using the translation model. This phase ensures that the final summaries are linguistically accurate thereby presenting information in the desired languages. By leveraging the translation capabilities of the model, we enhance the overall summarization process, allowing for broader accessibility and understanding of the content in the targeted linguistic contexts. This iterative approach of integrating translation back into the summarization workflow, contributes to appropriate summaries.

4.5. Performance Evaluation

The performance of the proposed methodology is quantitatively assessed using ROUGE and BERT scores, providing a comprehensive understanding of the summarization quality across multiple languages. The effectiveness of the model is validated through its ability to generate contextually accurate and linguistically appropriate summaries for diverse linguistic inputs.

This proposed methodology integrates the strengths of T5-base, translation models, and robust evaluation metrics, offering a comprehensive approach to multilingual summarization that aligns with the specific linguistic challenges posed by Indian English and Hindi

Model	Language	Rouge-1	Rouge-2	Rouge-L	Bert F1 Score
T5-base	English	0.2688	0.0845	0.1912	0.8090
T5-small	English	0.2515	0.0696	0.1271	0.7910
T5-base	Hindi	0.002	0.001	0.001	0.085
T5-small	Hindi	0.001	0.001	0.001	0.080

Table 1

Initial Performance Before Fine-Tuning the Models

Initially before fine-tuning the model with the FIRE 2023 dataset, the model performance was very poor and generated inaccurate summaries.

5. Results and Discussion

The results for the model can be comparatively analysed using table 1 and table 2. There is a much significant improvement in the performances of the model summarising capability on the translation, fine-tuning and back-translation.

Metric	ROUGE-1	ROUGE-2	ROUGE-L	BERT F1 Score
English	0.3022	0.1111	0.2504	0.8616
Hindi	0.2701	0.1214	0.2237	0.6782

Table 2

m2023-t5-base: ROUGE and BERT Scores for English and Hindi Summarization

The m2023-T5-base model performed well and helped in achieving rank 2 in the shared task of FIRE 2023 in Indian Language Summarisation.

6. Conclusion

In this comprehensive investigation into article summarization across Indian languages, encompassing Hindi and English, our study delved into the evaluation metrics, particularly focusing on ROUGE and BERT scores for English. The meticulous analysis of these scores provides valuable insights into the effectiveness of our summarization techniques.

For English, the ROUGE scores revealed commendable results: ROUGE-1 at 0.3022, ROUGE-2 at 0.1111, ROUGE-4 at 0.042, and ROUGE-L at 0.2504. These scores reflect the precision of our models in capturing unigram, bigram, and long-range dependencies, showcasing their proficiency in generating accurate and coherent summaries.

The BERT scores further substantiated the success of our summarization approach: Precision (P) at 0.8505, Recall (R) at 0.8733, and F1 Score at 0.8616. These scores emphasize the model's ability to comprehend and reproduce the semantic nuances present in the source articles, indicating a robust understanding of contextual information.

The positive outcomes in English summarization lay a strong foundation for extending these techniques to other Indian languages, addressing the linguistic diversity inherent in the dataset.

In conclusion, our study not only attests to the efficacy of our summarization models, as evidenced by the impressive ROUGE and BERT scores for English, but also sets the stage for further exploration and adaptation of these methodologies for Indian languages. The promising results underscore the potential of our techniques to enhance summarization across diverse linguistic landscapes, contributing to the advancement of natural language processing in the context of Indian languages.

7. Acknowledgements

We thank the FIRE 2023 organising committee for conducting this shared task for Indian Language Summarization.

References

- [1] Krishnakumar, A., Naushin, F., Mrithula, K.L. and Bharathi, B., 2022, December. Text summarization for Indian languages using pre-trained models. In Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation, Kolkata, India.
- [2] Mackie, S., McCreadie, R., Macdonald, C. and Ounis, I., 2014. Comparing algorithms for microblog summarisation. In Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings 5 (pp. 153-159). Springer International Publishing.
- [3] Singh, A.K., Varma, V. and Gupta, M., 2018. Neural approaches towards text summarization. International Institute of Information Technology Hyderabad.

- [4] Agrawal, A., Jain, R., Divanshi and Seeja, K.R., 2023, February. Text Summarisation Using BERT. In International Conference On Innovative Computing And Communication (pp. 229-242). Singapore: Springer Nature Singapore.
- [5] Paulus, R., Xiong, C. and Socher, R., 2017. A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.
- [6] Rahman, A., Rafiq, F.M., Saha, R. and Rafian, R., 2018. Bengali text summarization using TextRank, Fuzzy C-means and aggregated scoring techniques (Doctoral dissertation, BRAC University).
- [7] Kumari, K. and Kumari, R., 2022. An Extractive Approach for Automated Summarization of Indian Languages using Clustering Techniques. In Forum for Information Retrieval Evaluation (Working Notes)(FIRE). CEUR-WS. org.
- [8] Satapara, S., Modha, B., Modha, S. and Mehta, P., 2022, December. Fire 2022 ilsum track: Indian language summarization. In Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation (pp. 8-11).
- [9] Satapara, S., Modha, B., Modha, S. and Mehta, P., 2022. Findings of the first shared task on indian language summarization (ilsum): Approaches, challenges and the path ahead. Working Notes of FIRE, pp.9-13
- [10] Singh, S., Singh, J.P. and Deepak, A., 2022, December. Deep Learning based Abstractive Summarization for English Language. In Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation, Kolkata, India.
- [11] Agarwal, A., Naik, S. and Sonawane, S., 2022, December. Abstractive Text Summarization for Hindi Language using IndicBART. In Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation, Kolkata, India.
- [12] Urlana, A., Bhatt, S.M., Surange, N. and Shrivastava, M., 2023. Indian language summarization using pretrained sequence-to-sequence models. arXiv preprint arXiv:2303.14461.
- [13] Shrey Satapara, Parth Mehta, Sandip Modha, and Debasis Ganguly. *Indian Language Summarization at FIRE 2023*. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Goa, India. December 15-18, 2023*. ACM, 2023.
- [14] Shrey Satapara, Parth Mehta, Sandip Modha, and Debasis Ganguly. *Key Takeaways from the Second Shared Task on Indian Language Summarization (ILSUM 2023)*. In *Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, Goa, India. December 15-18, 2023*. Edited by Kripabandhu Ghosh, Thomas Mandl, Prasenjit Majumder, and Mandar Mitra. *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.