Hindi Text Summarisation

Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), 110020, Delhi, India

Abstract

This paper investigates Hindi text summarization using the ILSUM 2023 dataset, a recently released benchmark containing 21,225 new articles. We evaluate the performance of state-of-the-art multilingual transformer-based models on this task. Additionally, we explore knowledge distillation techniques, achieving successful implementation of self-distillation to enhance model performance. To improve data quality, we incorporate pre-processing steps such as transliteration and sentence truncation. Our work contributes to the advancement of Hindi text summarization by benchmarking advanced transformer architectures and demonstrating the effectiveness of self-distillation for performance improvement.

Key words. text summarisation, hindi, , handwriting recognition, RAG

1. INTRODUCTION

In the digital era, the proliferation of textual content in regional languages such as Hindi has highlighted the critical need for specialized natural language processing (NLP) tools. Hindi, one of the principal languages spoken globally, generates immense volumes of text online and offline every day. Effective summarization tools are crucial for individuals and organizations to quickly understand the gist of content without delving into the full text. This project seeks to bridge the gap in this domain by applying advanced computational methods to improve Hindi text summarization.

1.1. Motivation

The motivation of this work is real-world challenges and the need that arises from the overwhelming flow of Hindi text data over various platforms, including news portals, blogs, and official communication. Here are some reasons for the importance of this endeavor, each human-centered:

- **Easing Information Overload**: In an effort to keep abreast of the information in this world today, which can be cumbersome to journalists and researchers, automated summarization tools bring about a lot of help in a bid to have them well-informed and not overwhelmed
- **Enhancing Accessibility**: Long documents can be off-putting to many, particularly those within time-strapped environments or with low reading literacy. Summarisation can be the mechanism by which this barrier is overcome, allowing, for example, government policy or public health advisory information to be condensed into manageable, understandable sizes. This will also form an important and effective means of speedy and wider dissemination of important information, especially to the masses that speak Hindi.
- **Supporting Education**: Most of the times, with content turning digital and especially in Hindi-speaking belts, students and educators can make use of summarization tools pulling key information from volumes of material. This adds on to better understanding and also supports retention, making learning quite effective and interesting.
- **Empowering Businesses**: Businesses, which function in Hindi-speaking areas, need to plow through gigantic amounts of local consumer feedback, market reports, or competitor news, and be able to sift through them effectively and efficiently. Summarization tools can empower businesses operating in Hindi-speaking area derive powerful actionable insights from text data, thus ensuring faster and more informed decision-making.

2. Related Work

2.1. Indian Language Summarization using Pretrained Sequence-to-Sequence Models

Urlana et al. (2022) addresses the challenge of text summarization in underrepresented languages, specifically Hindi and Gujarati, alongside English, utilizing various pretrained sequence-to-sequence models. In the realm of natural language processing, sequence-to-sequence models such as PEGASUS, BART, T5, and ProphetNet have shown substantial capabilities in text summarization tasks, a field that has seen limited datasets of high quality, particularly for Indian languages. By leveraging multilingual datasets like XL-Sum and MassiveSumm, and fine-tuning models like MT5, MBart, and IndicBART, this study provides a comprehensive evaluation of model performance across different languages, showcasing significant improvements in summarization tasks. The research not only fills a gap by developing resources and methodologies for Indian language NLP but also

provides a comparative analysis that helps in understanding the strengths and limitations of various models. Through their methodological innovations, such as the application of k-fold cross-validation in scenarios of limited data, the authors contribute to advancing the field, setting a benchmark for future work in multilingual text summarization.

2.2. Noisy Self-Knowledge Distillation for Text Summarization

Liu et al. (2021) explores the application of self-knowledge distillation techniques to text summarization, aiming to address limitations posed by traditional maximum likelihood training methods on noisy and single-reference datasets. The authors propose a novel framework where a student model learns from a teacher model's softened output distributions, integrating multiple noise signals during training to model uncertainty more effectively. This approach not only helps regularize the training process but also enhances the robustness and performance of both pretrained and non-pretrained summarization systems across various benchmarks. By employing these techniques, the paper reports state-of-the-art results, demonstrating significant improvements over existing methods. This study contributes to the broader field by suggesting that incorporating noise and leveraging self-distillation can effectively mitigate the challenges of overfitting and data quality in neural text summarization models .

2.3. Distillation Knowledge applied on Pegasus for Summarization

Niccolai (2024) focuses on exploring knowledge distillation applied to the Pegasus model for text summarization within the domain of Natural Language Processing (NLP). The study centers on reducing the complexity of the large, state-of-the-art Pegasus model, a transformer-based architecture, by utilizing knowledge distillation techniques. This process involves compressing the model into a smaller size without significantly sacrificing performance. The research assesses various distilled models by evaluating their performance metrics, emphasizing the importance of deploying such technologies in devices with limited computational resources and inconsistent internet connectivity. By experimenting with different sizes of distilled models, Niccolai aims to understand the trade-offs between model size, computational efficiency, and performance in text summarization tasks .

3. Methodology

3.1. Pre-processing

- **Transliteration**: Using AI4BHĀRAT, we transliterated all the English text which is present in the columns "Article", "Heading" and "Summary". We utilised re.findall() function to find all the possible English alphabets in the documents. We also kept the hyper parameter of beam_width as 15, since it provided us more accurate inferences.

- **Sentence Trunctation**: Longer articles were truncated to manageable sizes to prevent model overload and to focus on the most relevant information, which helps in maintaining computational efficiency and model performance.

3.2. Fine Tuning

3.2.1. IndicBart

IndicBART is a multilingual, sequence-to-sequence pre-trained model focusing on Indic languages and English. It currently supports 11 Indian languages and is based on the mBART architecture.

- **Training on Summarization**: The model was finetuned exclusively on the task of summarization. The training involved feeding the model with pairs of full articles and their corresponding summaries, allowing the model to learn the pattern of condensing text into a summary effectively.

- **Hyperparameter Optimization**: Careful adjustments were made to the learning rate, batch size, and the number of epochs to find the optimal balance for training convergence. Typically, a lower learning rate and a moderate batch size were employed to ensure stable learning without overfitting.

- **Regularization Techniques**: Techniques like dropout and attention masking were used to enhance generalization and prevent the model from memorizing specific textual patterns from the training data.

3.2.2. mT5-multilingual-XLSum

Fine-tuned mT5 model tuned on XL-Sum dataset. Training on Summarisation: Hugging face docs were used to code the summarization

Model Selection We selected the mT5 (multilingual T5) model, specifically the variant pre-trained on the XLSum dataset, due to its robust multilingual capabilities and its proven effectiveness in summarization tasks across various languages. The mT5-multilingual-XLSum model, being pre-trained on a diverse set of news articles, offers a strong foundation for understanding and generating news summaries, which aligns well with the nature of the ILSUM 2023 dataset.

Fine-tuning Process The fine-tuning was conducted on a filtered subset of the ILSUM 2023 dataset, which was specifically curated to include a balanced mix of article lengths and topics to provide comprehensive coverage of the linguistic features in Hindi. The mT5 model was fine-tuned using a custom learning rate scheduler to adaptively adjust the learning rate based on training progress, optimizing the convergence speed and final model performance. The AdamW optimizer was employed for its effectiveness in handling sparse gradients and for better weight regularization.

3.3. Self-Distillation

Self distillation on IndicBart Attempt at Cross distillation on IndicBart. We initialised two instances of IndicBART. The first model acts as the "Teacher Model," and the second as the "Student Model. Utilize the IndicBART tokenizer to process text data. This involves defining prefix tokens (like 'summarize:') to signal the summarization task. We used a data collator as well so that the embedding that the tokenizer gave after mapping was ready for training.

This loss measures the discrepancy between the predicted probability distribution generated by the student model and the actual distribution (ground truth labels). It is a standard loss used in classification tasks, including text summarization, to directly encourage the student model to predict the correct classes (or tokens, in the case of summarization). The formula for cross-entropy loss is:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

3.3.1. KL-Divergence Loss (for Distillation)

KL-Divergence is used here as a measure of how one probability distribution (the output from the teacher model) diverges from a second, expected probability distribution (the output from the student model). This is essential for distillation, as it encourages the student to mimic the "soft" probabilities of the teacher model, beyond merely predicting the correct label. The soft probabilities can convey additional information about the confidence of various predictions. The formula for KL-divergence is:

$$\mathrm{DKL}(P\|Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

3.Combined Loss Calculation

Finally, the overall loss used to update the student model is a weighted combination of the cross-entropy loss and the distillation loss. This combination allows the model not only to learn directly from the true labels but also to benefit from the additional information contained in the teacher model's predictions:

3.4. Crossdistillation

Cross-Entropy Loss: Cross distillation is a technique used in machine learning where knowledge is transferred between two or more models to mutually improve each other's performance. Unlike traditional knowledge distillation, which typically involves a one-way transfer from a more knowledgeable teacher model to a less knowledgeable student model, cross distillation allows for a bidirectional or multidirectional exchange of information. This process can occur between models of similar complexities or between models trained on different tasks or data subsets.

3.4.1. 1. Cross-Entropy Loss

This is used to measure the discrepancy between the predicted labels by the student model and the actual ground truth labels. The formula is:

Where y is a binary indicator (0 or 1) if class label c is the correct classification for observation o , and p is the predicted probability observation o is of class c.

3.4.2. 2. KL-Divergence Loss

KL-Divergence is used here for distillation to measure how the student's probability distribution deviates from the teacher's. It's used to make the student's predictions closer to the soft probabilities provided by the teacher.

4. Analysis and Observation

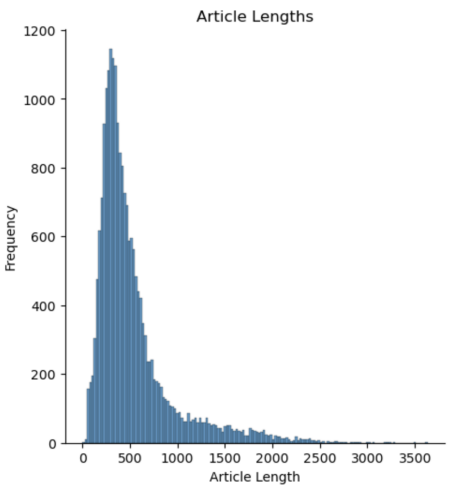

Fig. 1: For Train Da

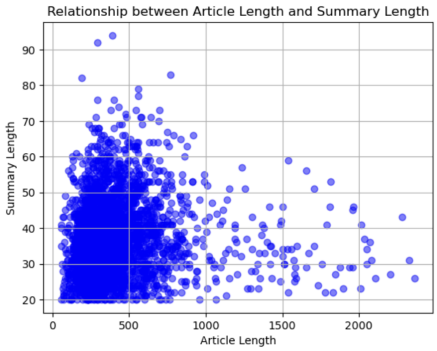

Fig. 2: For Test Data

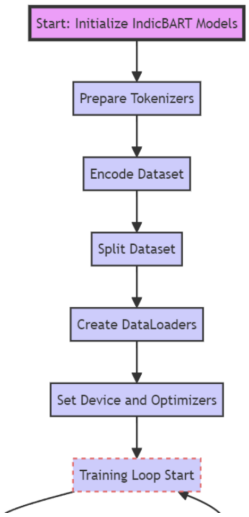

Fig. 3: Part 1

	Rouge-1	Rouge-2	Rouge-L
Self Distillation on IndicBart (On first Training)	0.5065	0.3984	0.4737
Self Distillation on IndicBart (After Learning from the Parent Model)	0.5876	0.4944	0.4944
Fine Tuning (Indic Bard)	0.5612	0.4512	0.5161
Fine Tuning (mT5-XL)	0.5378	0.4263	0.4987

Table 1: Results

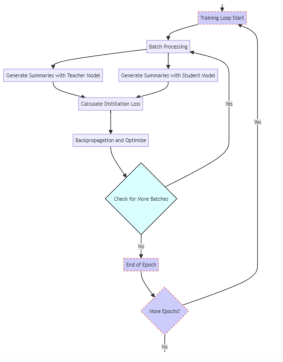

Fig. 4: Part 2

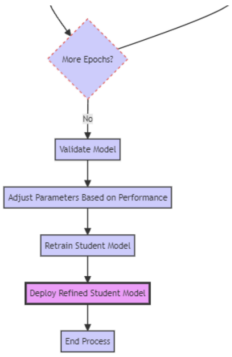

Fig. 5: Part 3

5. Conclusion and Future Work

The results of implementing self-distillation on IndicBART for Hindi text summarization clearly indicate its efficacy, as demonstrated by the improvement in ROUGE scores. Initially, the ROUGE-1, ROUGE-2, and ROUGE-L scores for the first training phase of self-distillation on IndicBART were 0.5065, 0.3984, and 0.4737, respectively. After learning from the parent model, these scores significantly improved to 0.5876, 0.4944, and 0.4944, highlighting the advantage of the self-distillation process. This approach not only surpassed the initial training but also outperformed the results of straightforward fine-tuning of IndicBART and mT5-XL models. The enhanced performance after self-distillation underscores the effectiveness of using advanced training techniques like self-distillation, which leverages the nuanced understanding and implicit knowledge encoded by the parent model to boost the summarization capabilities of the student model. This suggests that self-distillation can be a valuable strategy for enhancing neural network performance, particularly in complex NLP tasks like text summarization.

REFERENCES

Liu, F. N. et al. 2021, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics,)

Niccolai, L. 2024, Knowledge Distillation on Pegasus Model for Text Summarization

Urlana, A., Bhatt, S. M., Surange, N., and Shrivastava, M. 2022, Forum for Information Retrieval Evaluation,)