Text Summarization for Hindi Language

Unlocking Clarity, One Summary at a Time

GitHub Repo

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

Authors

Nikhil Suri (2021268)
Maanas Gaur (2021537)
Lakshya Aggrawal (2021535)
Lakshya Kumar (2021536)

Affiliations

IIIT - Delhi,
CSE556 - Natural Language Processing
Prof. Md Shad Akhtar

01 Introduction

Text summarization in Hindi is challenging due to its rich morphological structure & diverse linguistic nuances

Hindi Text Example:

"भारतीय राजनीति में उठापटक एक आम बात है। हाल ही में, विभिन्न राज्यों में चुनावी माहौल ने जनता की राय और सरकार की नीतियों के बीच गहरे मतभेदों को उजागर किया है। इन चुनावों के परिणाम ने न केवल राजनीतिक दलों की रणनीति पर प्रभाव डाला है, बल्कि आम आदमी की दैनिक जीवन और अर्थव्यवस्था पर भी गहरा असर पड़ा है। आर्थिक मंदी, नौकरी की कमी, और सामाजिक अशांति के बीच, यह महत्वपूर्ण है कि सरकार नई नीतियाँ लागू करे जो जनता के हित में हों।"

02 Objective

- To develop NLP models that accurately generate summaries of Hindi texts
- To enhance readability and comprehension for different use-cases such as real-time news aggregation

03 Methodology

Tokenization

- To convert text into a format suitable for model input, ensuring efficient handling of Hindi text

Model Approach

- Fine-tuned IndicBART and MT5, & a custom Transformer model to accommodate the syntactic and semantic intricacies of Hindi.

04 Model Description

Encoder

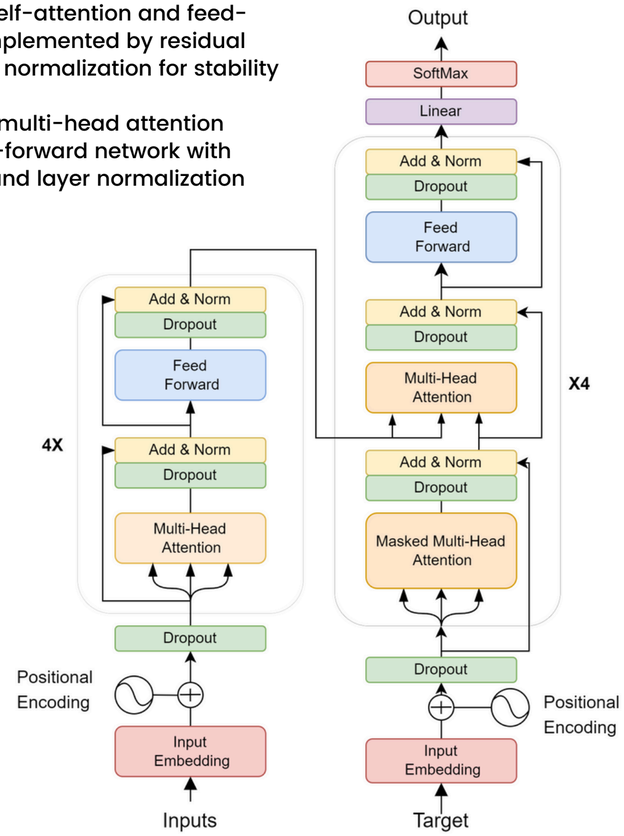
- Layers incorporating self-attention and feed-forward networks, complemented by residual connections and layer normalization

Decoder

- Layers that have dual multi-head attention mechanisms & a feed-forward network with residual connections and layer normalization

Pre-trained Models

- IndicBART is a multilingual seq2seq model, features 6 encoder-decoder layers, 16 attention heads totaling 244M parameters
- MT5 is a multilingual text-to-text model with 12 encoder-decoder layers and 12 attention heads each



05 Results/Findings

Model	BERTScore Precision	BERTScore Recall	BERTScore F1
IndicBART	0.789	0.852	0.819
Transformer Model	0.612	0.542	0.737
MT5 Model	0.783	0.765	0.773

QUALITATIVE ANALYSIS: T MODEL

- Gen: बहुजन पार्टी की राष्ट्रीय पार्टी की अध्यक्ष मायावती ने सत्ताईस सितंबर को किसान संगठनों द्वारा बंद बंद का समर्थन किया है
- Ref: बहुजन समाज पार्टी की अध्यक्ष मायावती ने कल सत्ताईस सितंबर को किसान संगठनों द्वारा बुलाए गए भारत बंद का समर्थन किया है

06 Conclusion

This project highlights IndicBART's superior performance in summarizing Hindi texts, outperforming both the Transformer and MT5 models. The results affirm the strength of fine-tuned, pre-trained models in tackling specific linguistic challenges in NLP

FUTURE WORK

- Improving the model's handling of Hindi's linguistic nuances
- Expanding the dataset to encompass a broader spectrum of text genres for a more comprehensive performance evaluation

Refrences

- Saini, Maitree et al. 2019. "Unsupervised Abstractive Summarization of Bengali Text Documents." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.
- Liu, Yinhan et al. 2020. "Fine-tuning & Decoding in Multilingual Denoising Pre-training for Neural Machine Translation." 2020. ArXiv.
- Tang, Gongbo et al. 2020. "Multilingual Denoising Pre-training for Neural Machine Translation." 2020. ArXiv.
- Wubben, Sander. 2021. "Text Summarization: A Low Resource Challenge." Transactions of the Association for Computational Linguistics.
- Mandl, Thomas et al. 2021. "Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages." FIRE 2021.

Acknowledgement

- Special thanks to Prof. Shad Akhtar, whose guidance significantly contributed to the completion of this project.