Text Summarisation for Indian Languages (Hindi)

Anurag Gupta (2021451), Manshaa Kapoor (2021540), Rajat Raghav (2020568), Tanuj Khatri (2020578)

INTRODUCTION

Text summarization is a crucial task in natural language processing (NLP) that aims to condense a given piece of text while retaining its key information. With the increasing availability of digital content in Indian languages, there is a growing need for effective text summarization methods tailored to these languages. The ILT-2023 dataset presents a unique challenge due to the presence of code-mixing and script mixing, making it distinct from previous datasets used in other languages. In this project, we focus on Task 1 of the ILT-2023 dataset, which involves generating meaningful fixed-length summaries for articles in Indian languages. The dataset provides articles in Hindi, Gujarati, and Bengali, offering a diverse linguistic landscape for research and development in text summarization. We will mainly focus on Hindi Language. We employ both extractive and abstractive summarization approaches to generate summaries for the given articles. We evaluate our models using standard metrics such as ROUGE 1,2 and 4 F1 and BERT-Score, which are commonly used in summarization tasks to assess the quality of generated summaries. Overall, our goal is to explore effective text summarization techniques for Indian languages, particularly addressing the challenges posed by code-mixing and script mixing, and to contribute towards advancing research in this important area of NLP.

METHODOLOGY

Data Preprocessing -

- **Tokenization:** We tokenize the input text into words or subwords using the tokenizer specific to each model (e.g., WordPiece tokenizer for BERT).
- **Data Cleaning:** We remove any unnecessary characters, symbols, and special characters from the text.

Models-

1. Bi-LSTM
2. BART
3. BERT

Custom Attention Mechanism-

- In all our models we are using a custom attention mechanism. This mechanism allows us to incorporate the article headings into the models. It calculates attention weights for the heading and article encodings and combines them to produce the final combined encodings.

Model Architecture -

- **Bi-LSTM :** Bidirectional LSTMs are a type of recurrent neural network (RNN) that processes the input sequence in both forward and backward directions, allowing the model to capture contextual information from both past and future tokens. In the context of text summarization, Bi-LSTMs are effective at understanding the context of a word or phrase within the input text, which is crucial for generating accurate and coherent summaries.
- **BART :** BART is a transformer-based model that generates output tokens sequentially, combining bidirectional processing and transformers' auto-regressive trait. Pre-trained using denoising autoencoding targets on a large text corpus, it provides useful summaries. BART is ideal for abstractive summary jobs due to its design and pre-training goals, enabling rearranging and rewriting input text.
- **BERT :** BERT is another transformer-based model that is pre-trained on a masked language modeling objective, where it learns to predict masked words in a sentence based on their context. BERT's bidirectional nature allows it to capture deep contextual relationships between words in the input text, which can be beneficial for understanding the meaning of the text and generating informative summaries.

Findings Comparision

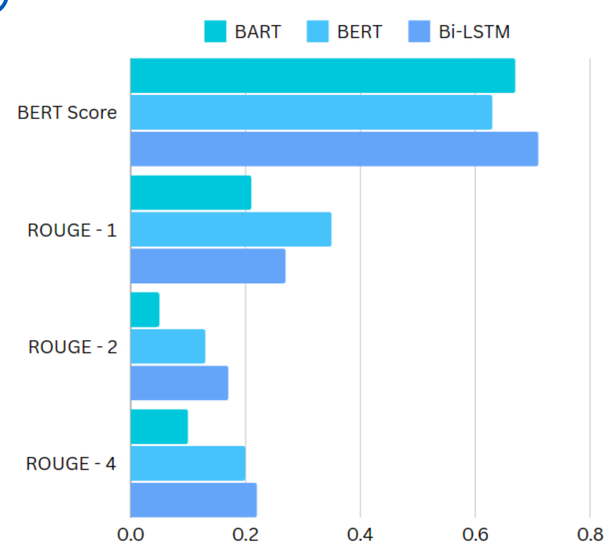

Figure A:Findings Comparision

EVALUATION METRICS :

- **Rouge Scores :** The ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-4) measure the overlap between the generated summaries and the reference summaries in terms of unigram, bigram, and 4-gram overlap, respectively. A higher ROUGE score indicates better agreement between the generated and reference summaries.
- **BERT Score :** The BERTScore measures the similarity between the model-generated summaries and human-written reference summaries using contextual embeddings from BERT. A higher BERTScore indicates a higher degree of similarity between the model-generated summaries and the reference summaries.

RESULTS :

- BERT outperforms both BART and Bi-LSTM in terms of ROUGE-1 and ROUGE-4 scores while Bi-LSTM achieves the highest ROUGE-2 score. This shows that Bi-LSTM and BERT are showing better agreement between the generated summaries and the reference summaries in terms of unigram, bigram, and 4-gram overlap than BART.
- Bi-LSTM achieves the highest BERTScore. This suggests that while Bi-LSTM may not perform as well in terms of n-gram overlap, it generates summaries that are more similar to human-written summaries according to the BERTScore metric.
- Using Attention Mechanism to incorporate headings gave us a boost in metrics of around 5-10 percent.

Conclusion :

In this study, we investigated the effectiveness of transformer-based models for text summarization tasks in Hindi, focusing on the BERT and BART and Bi-LSTM architectures. Our experiments involved training the models on a dataset of Hindi news articles and their corresponding summaries, using custom attention mechanisms to enhance the model's ability to generate concise and informative summaries by including headings from the dataset. The results showed that the BERT mode and Bi-LSTMl, outperformed the BERT model in Rouge F1 scores while BI-LSTM outperformed both of them in BERT score. Overall, our findings suggest that models augmented with custom attention mechanisms to incorporate headings, are highly effective for text summarization in Hindi and hold promise for further advancements in natural language processing tasks for Indian languages.