Problem Statement-2 Report

Research Article Summarization Using Advanced NLP Techniques

Completed on **Mar 24, 2025**

Team: Supreme BEings

- **Team Members**:
  1. Abhirup Saha
  2. Chandrachur Pramanik
  3. Joydip Paul
  4. Jyotirmoy Dutta
  5. Swarnava Banerjee

1. Table of Contents :

2. Introduction

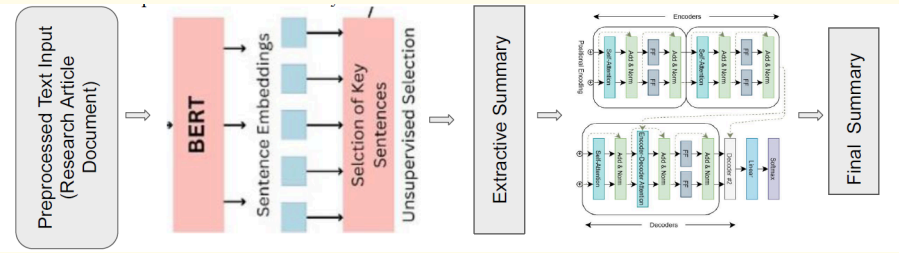
Dataset used : PubMed Dataset and **arXiv** Dataset for training the hybrid Model and
CompScholar DataSet for evaluation.

3. Data Preprocessing

- The **Data Preprocessing Pipeline** provides support for normal text and document(s).
- First the text received is cleaned, the leading and trailing white spaces are trimmed and the new line characters are replaced with a space to maintain sentence boundaries.
- To handle large files the processed text is divided into overlapping chunks of size compatible for transformer input, the overlapping helps in maintaining the context between the chunks.
- For this reason **SciBert tokenizer** has been used to tokenize the text and then the input text is divided into chunks based on token count.

4. Model Description

- An hybrid model is build combining the strength of **Extractive** and **Abstractive** methods of
summarization which can handle multiple research documents and along with the long documents
maintaining efficiency.
- The model accepts the preprocessed document and generates an extractive summary by identifying the key sentences using **fine-tuned BERT.**
- The extractive summary is then fed to the **fine-tuned T5** model to generate the abstractive summary from it. It provides the **Final Summary.**



5. Conclusion

“

1. 66We have developed the working prototype or to be specific the basic idea & code for the model.
2. Due to having no access to a dedicated GPU server, we could not execute this file.
3. Hope, judging team will consider this scenario.
4. Our **innovative approach** was to deploy an hybrid model so that extractive & abstractive records could be handled.

”

6. References

I. Chaudhari, N., Vora, D., Kadam, P., Vaishali Khairnar, Patil, S. and Kotecha, K. (2024). Towards efficient knowledge extraction: Natural language processing-based summarization of research paper introductions. IAES International Journal of Artificial Intelligence, [online] 14(1), pp.680–680. doi:https://doi.org/10.11591/ijai.v14.i1.pp680-691.

II. Kolambkar, V., Shingade, B., Matha, Y., Kasar, S. and Palve, P., 2024. A Survey of Text Summarization Using NLP. A Survey of Text Summarization Using NLP (March 24, 2024).