Topic: Bangla Text Summarization

Roll: 1907050

Paper 1:

“Bengali Text Summarization using TextRank, Fuzzy C-Means and Aggregate Scoring methods.”

<https://ieeexplore.ieee.org/document/8971039>

Authors: A. Rahman, F. M. Rafiq, R. Saha, R. Rafian and H. Arif

Conference: 2019 IEEE Region 10 Symposium (TENSYMP)

Publisher: IEEE

Summary:

To count them consistently, words are converted into their root form, the process is called Stemming. Stemming helps by converting all the variations like "run", "running", and "ran" to their basic form, like "run". Then low significant common words are removed. Term frequency (TF) measures how often a term appears in the document reflecting significance. Rare words are weighted higher. Numerical value-based scoring technique recognizes sentences that contain numerical data, such as statistics, figures, or measurements, as potentially more informative or relevant and gives higher priority. Sometimes longer sentences have more important information. Certain words act as cues or indicators of important information to give a score increment. For example, words like "মূল", "প্রধান", "প্রথম" (meaning "main", "primary", "first") may signal important content. The sentences in the first and last 10% of the paragraph are given a higher sentence position score. FCM algorithm clusters sentences into high and low importance clusters, from them the prior ones are taken for summary. TextRank ranks the sentences based on their similarity to other sentences in the text, top ranked ones form the summary. Aggregating all scoring method, top scored sentences form the summary. They compared their summary using ROUGE metric. FCM yields higher numbers of common sentences, indicating higher accuracy in generating summaries compared to TextRank and Aggregate Scoring.

Paper 2:

"Bengali News Abstractive Summarization: T5 Transformer and Hybrid Approach"

<https://ieeexplore.ieee.org/document/10410939>

Authors: K. M. Hasib, M. A. Rahman, M. I. Masum, F. D. Boer, S. Azam and A. Karim

Conference: 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)

Publisher: IEEE

Summary:

Their used dataset is taken from Kaggle containing news articles, each article be divided into three columns: category, summary, text. The T5 model consists encoder and decoder module. Encoder embeds input texts into vectors and applies self-attention mechanism to understand dependencies among words. Decoder generates output summaries based on this. The BenSumm model simplifies text by first breaking it into smaller pieces, removing unnecessary words, and tagging the parts of speech. Then, it groups similar sentences together using a method called hierarchical agglomerative clustering. After that, it creates a word graph to find important phrases in the text. Finally, it selects the most crucial sentences from each group to construct a summary that captures the main ideas of the original text. The dataset is cleaned by removing English words and duplicates. T5 model is fine-tuned using the preprocessed dataset. BenSumm performs unsupervised extractive summarization and thus doesn’t require training. For evaluation, ROUGE and BLEU were used to measure overlap with reference summaries and assesses text similarity respectively. Result indicates T5’s better success in generating summaries close to reference ones.

Paper 3:

"Pointer over Attention: An Improved Bangla Text Summarization Approach Using Hybrid Pointer Generator Network"

<https://ieeexplore.ieee.org/document/9689852>

Authors: N. Dhar, G. Saha, P. Bhattacharjee, A. Mallick and M. S. Islam

Conference: 2021 24th International Conference on Computer and Information Technology (ICCIT)

Publisher: IEEE

Summary:

They used neural network that can understand and generate text. It was designed to be able to create new words and copy existing words from original text. Their feature made sure to include all the important information from original text is copied without error and to avoid repeating words unnecessarily. Words can be taken from a listed vocabulary or can be copied from the original text to preserve the meaning, even if it were not in the list. To avoid repetition of same words too much, the used coverage mechanism that made it to pay attention to different parts of the original text and not to focus on the same thing again and again. For evaluation, two methods were used. Firstly, they compared the summary with human written summaries using a scoring system called ROUGE. Next, the feedback from people after reading their summaries. In both cases, they got better response for generating accurate and readable summaries.