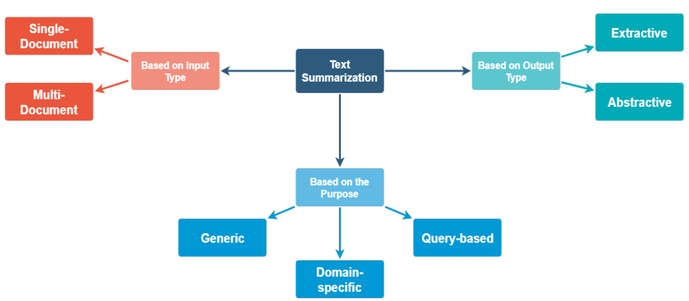
Text Summarization: Extractive and Abstractive Methods

◆Prabin Raj Shrestha ◆ Shivani Agarwal ◆ Neha Gaddam ❖Under Prof. Lu Xiao

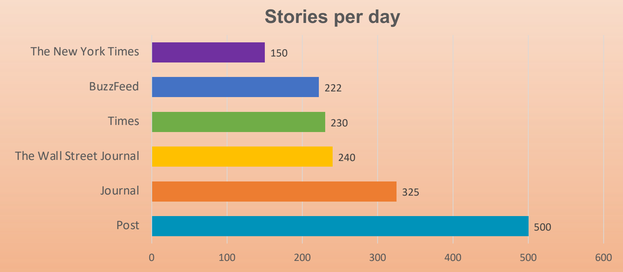
Introduction

Text Summarization is the process of creating a summary of a certain document that contains the most important information of the original one, the purpose of it is to get a summary of the main points of the document.
Text summarization has become increasingly important due to the vast amount of data available today.
Extractive and abstractive are the main approaches used to create summaries. Extractive summarization has matured, and research has shifted towards abstractive and real-time summarization.



Problem Statement

- The demand for condensed and focused article summaries is increasing due to the vast volume of daily published articles and limited time to process them.
- Due to the overwhelming volume of data, manual summarization is not feasible. As a result, automatic methods and algorithms are required to generate summaries efficiently and effectively.
- There are challenges in identifying relevant content, avoiding redundancy, maintaining coherence, and preserving the tone and context of the original text.



According to research, there are **2 to 3 million** news articles that make it to both offline and online platforms every 24 hours

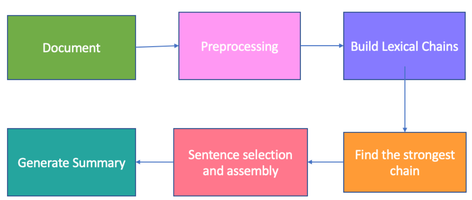
Source: How Many News Articles are Published Every Day in 2023? https://earthweb.com/how-many-news-articles-are-published-every-day/

Extractive Summarization

Extractive summarization is a way of summarizing text by selecting and combining the most important sentences or phrases from the original text. This method keeps the original wording and order of the selected sentences. It is simpler than abstractive summarization because it doesn't require generating new sentences. However, the summaries produced may be less coherent and less readable than abstractive summaries.

Approaches

Corpus based approach:
It leverages the structural distribution of words using internal or external corpus for summarization. (e.g., WordNet)
Cohesion based approach:
It considers cohesive relations between the concepts within the text such as (antonyms, repetitions, synonyms etc.) using Lexical Chains.



Graph based approach:
Each sentence in the text is represented as a vertex and a graph is constructed around all the sentences, where the edges correspond to the interconnections between the sentences. LexRank and TextRank are two such techniques

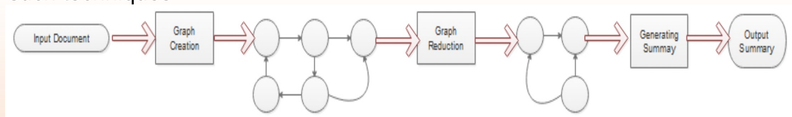
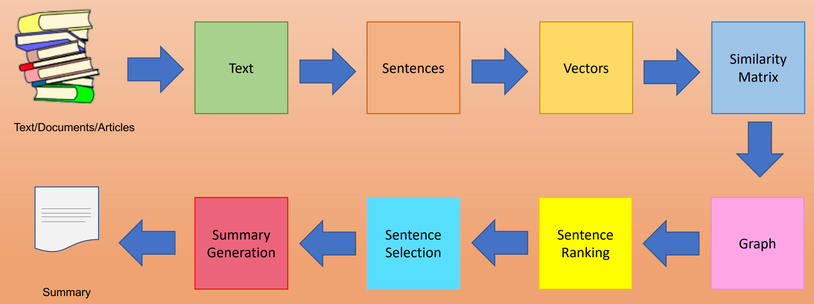


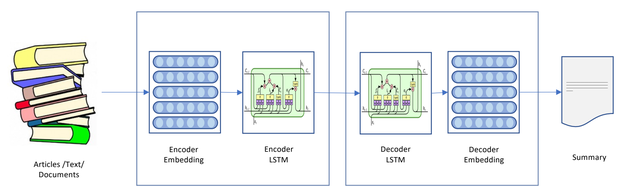
Fig. Graph reduction

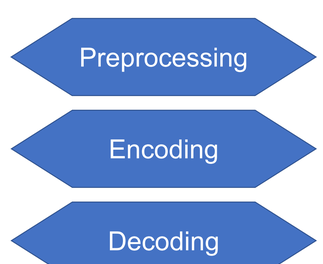
Machine Learning Based approach:
Document summarization can use supervised or semi-supervised learning. In supervised learning, sentences are labeled based on clues or manual tagging, and a binary classifier is trained to assign a summary likelihood score to each sentence.



Abstractive Summarization

Abstractive summarization creates a shorter summary of text by generating new sentences that convey the same meaning as the original text. It uses natural language generation techniques to produce high-quality summaries that retain essential information while maintaining readability. It is particularly useful for large amounts of data, but more challenging than extractive summarization. Abstractive summarization of multi-documents aims to distill the most important ideas while keeping the summary concise and coherent.





The input text undergoes preprocessing to eliminate unwanted elements and tokenize it into words or subwords.

An encoder-decoder architecture using seq2seq approach is employed to process the preprocessed input text, where the encoder generates a fixed-length embedding of the text.

The decoder predicts the next probable word using the encoder's fixed-length representation to generate a summary. It is trained to produce a coherent and readable summary that summarizes the critical information in the input text.

The model is trained on a paired input-output dataset using cross-entropy loss function to minimize the difference between the predicted and target summaries.

The quality of generated summaries is assessed using ROUGE metric that compares them to reference summaries based on shared n-grams.

The model can be improved on specific types of texts or topics by fine-tuning it on domain-specific data through transfer learning techniques.

The final trained model can be deployed as an API or integrated into a larger application to automatically generate summaries of input text.

Conclusion

The exponential growth of digital data has made it difficult to navigate and comprehend vast amounts of information, which has led to a growing need for automatic text summarization. This task can be achieved through extractive or abstractive summarization techniques, both of which have made significant progress through natural language processing and machine learning. However, further research and development are necessary to improve the accuracy and effectiveness of these algorithms, as the importance of summarization is expected to increase with the continued growth of digital data.

References

- Review of automatic text summarization techniques & methods https://tinyurl.com/S1319157820303712
- Neural Abstractive Text Summarization with Sequence-to-Sequence Models https://dl.acm.org/doi/abs/10.1145/3419106
- Text Summarization: A Brief Review https://tinyurl.com/978-3-030-34614
- Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks https://arxiv.org/pdf/1807.08000.pdf
- Bootstrap Your Text Summarization Solution with the Latest Release from NLP-Recipes https://tinyurl.com/ba-p1268809