Report – Text Summarization

Introduction

The problem of Text summarization in the domain of Natural Language Processing, refers to the task of condensing a given piece of text into a short summary. It is a strenuous problem of Natural Language Processing (NLP) due to difficulty in interpreting every point of the text in a document. This requires a precise analysis of the text in various steps such as semantic analysis, lexical relations, named entity recognition, etc., which can be accomplished with a great deal of word knowledge only. Since it is hard to obtain the word knowledge in various aspects such as meaning of a word with respect to other content, related words, inferential interpretation, sentence generation, etc., generating abstracts as summaries have become complex. This type of summarization is classified as an abstractive summarization in NLP.

However, an approximation, which is classified as extractive summarization, is more flexible. In particular, the system is required to identify the most relevant/ significant contents of the text, extract them, order them, and return them to the user. Although extractive summarization tasks have been a popular research topic since 1958 (Luhn, 1958), yet it is a great challenge to summarize a text automatically using a computational system like a human generated summary. Several aspects about a good summary have been introduced by researchers.

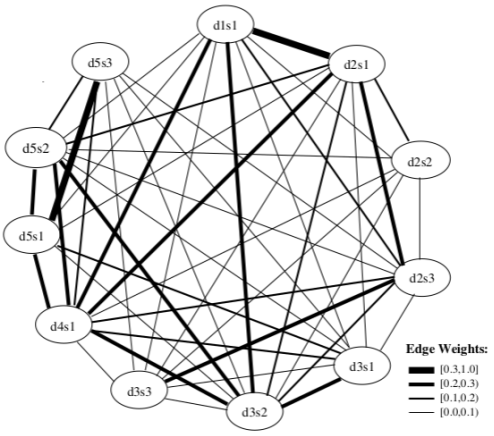
Literature Study

Text Summarization methods are publicly restricted into abstractive and extractive summarization, third type - a hybrid of two is also being focussed in recent years as Intuition supports that we will be able to surpass the individual performance of methods if we combine the best of two methods.

Extractive Text Summarization methods can be broadly classified as Unsupervised Learning and Supervised learning methods, recent works rely on Unsupervised Learning methods for text summarization. An extractive summarization technique consists of selecting vital sentences, paragraphs, etc., from the original manuscript and concatenating them into a shorter form. The significance of sentences is strongly based on statistical and linguistic features of sentences. There are various kinds of extractive summarization methods such as - concept based, statistical method, fuzzy logic based approach, graph based approach etc. We used the following two Extractive Text summarization Techniques:-

Statistical Methods*:* Such methods use statistical features of the document to identify the important pieces of the text. In a statistical method, a sentence is selected based on features like word frequency, position of the sentence, indicator phrases, title, location, and other features regardless of the meaning of the sentence. The method calculates the scores of the selected sentences and chooses a few highest scoring sentences to create the summary.

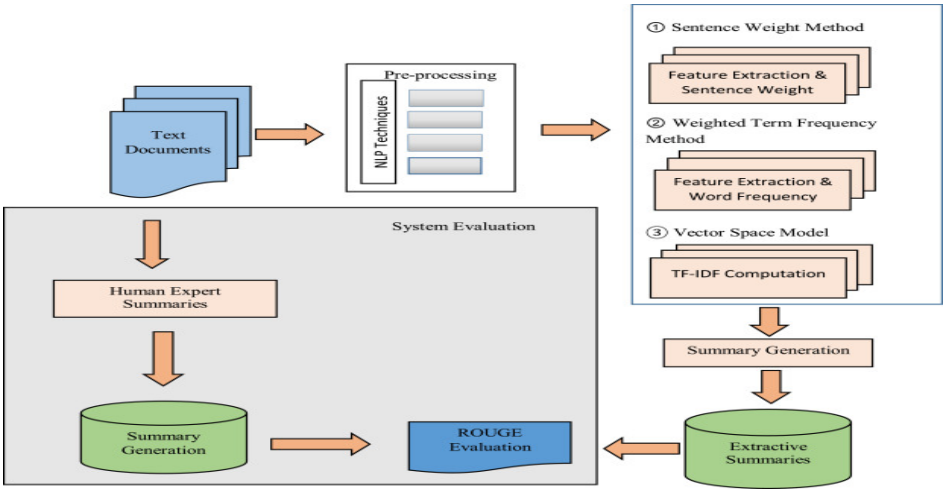
Graph Based Methods: Graph-based models are extensively used in document summarization since graphs can efficiently represent the document structure. Extractive text summarization using external knowledge from Wikipedia incorporating a bipartite graph framework has been used. One such graph based approach is LexRank in which the salience of the sentence is determined by the concept of Eigenvector centrality. The sentences in the document are represented as a graph and the edges between the sentences represent weighted cosine similarity values. The sentences are clustered into groups based on their similarity measures and then the sentences are ranked based on their LexRank scores similar to the PageRank algorithm except that the similarity graph is undirected in LexRank method.



weighted cosine similarity graph

We have implemented the following method for abstractive summarization -

Abstractive Method - Transformers: Before transformers, most state-of-the-art NLP systems relied on gated RNNs, such as LSTMs and gated recurrent units (GRUs), with added attention mechanisms. Transformers also make use of attention mechanisms but, unlike RNNs, do not have a recurrent structure. This means that provided with enough training data, attention mechanisms alone can match the performance of RNNs with attention.



Text summarization Pipeline

Central Idea

Our central idea is to present a comparative study between extractive and abstractive method of summarization, namely statistical, lexrank method for extractive and Transformer (BART) for abstractive summarization.

Experimental Setup

Primary Platform - Google Colab
Libraries - lexrank, transformers, sumy, nltk, spacy, pytorch

Conclusions

Through the comparison between the models, evaluation metrics and summary generated we can clearly see that abstractive summarization is in general more close to a summary generated by a human than extractive summarization, it was logical to assume the same as abstractive methods try to form new sentences based on internal meanings of words unlike the extractive summarization methods. However, the results are still far unlike quality human summaries even though many techniques have been proposed and researched upon.

References

1. https://ieeexplore.ieee.org/document/9458057 - Recent Progress on Text Summarization
2. https://www.bhu.ac.in/research_pub/jsr/Volumes/JSR_64_01_2020/48.pdf - A Review on Text Summarization Techniques
3. https://arxiv.org/abs/1910.13461 - BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension
4. https://www.jair.org/index.php/jair/article/view/10396 - LexRank: Graph-based Lexical Centrality as Salience in Text Summarization.
5. https://en.wikipedia.org - wikipedia
6. Numerous medium articles related to the topics.