NATURAL LANGUAGE PROCESSING

COMPARATIVE STUDY OF EXTRACTIVE SUMMARIZATION TECHNIQUES

JANAM VAIDYA (201701027) AND RUSHABH GAJAB (201701432)

Introduction

There are two types of summarization: abstractive and extractive summarization. Abstractive summarization basically means rewriting key points while extractive summarization generates summary by copying directly the most important sentences from a document.

In this literature we aim at comparing five of the extractive summarizers of Natural Language Processing. We compare these on parameters such as the Recall score, Precision score and F1-score. In conclusion,we would like to order these techniques according to the F1-score that is obtained.

Techniques used and their details:

1. **LexRank-** It is an unsupervised way of text summarization, that is based on the graph-based centrality scoring of sentences. The summarization is based on the similarity of the sentences. Thus, if one sentence is very similar to many others, it will likely be a sentence of great importance.To get highly ranked and placed in the summary. A sentence must be similar to many sentences that are in turn also similar to many other sentences

   Each document is treated as a graph and each sentence is treated as a node, whereas the edges represent cosine similarity between sentences. These nodes are then ranked on the basis of the PageRank algorithm, in which the importance of a node is decided iteratively by looking at the other sentences connected to the current sentence.[2]

2. **BERT Extractive Summarizer-**This works by first embedding the sentences, then running a clustering algorithm, finding the sentences that are closest to the cluster's centroids. The basis for this algorithm is the unsupervised TextRank model.

   The BERT summarizer has 2 parts: a BERT encoder and a summarization classifier.The task of extractive summarization is a binary classification problem at the sentence level. We want to assign each sentence a label $y_i \in \{0,1\}$ indicating whether the sentence should be included in the final summary. Therefore, we need to add a token before each sentence. After we run a forward pass through the encoder, the last hidden layer of these tokens will be used as the representations for our sentences.[1]

3. **TextRank Summarizer-** TextRank is an unsupervised text summarization technique. Here, all the text contained in the articles is concatenated and then split into individual sentences. Then we find the vector representation of the for all the sentences. Similarities between sentence vectors are then calculated and stored in a matrix. The similarity matrix is then converted into a graph, with sentences as vertices and similarity scores as edges, for sentence rank calculation.Finally, a certain number of top-ranked sentences form the final summary.
4. **Luhn Summarizer-** Luhn's method is a simple technique in order to generate a summary from given words. ... Luhn states that this is first done by doing a frequency analysis, then finding words which are significant, but not unimportant English words.[4] Out of the given summarization techniques this is the least effective.
5. **LSA Summarizer-** In this approach the document is represented as a term-sentence matrix which is then projected into the semantic space by using singular value decomposition . The sentences corresponding to top-k singular values are then included in the summary[3].

Results

Summarization Technique	Recall	Precision	F1-score
LexRank	0.26041666	5.0	0.78125
BERT	0.577777777777	5.2	1.7333
TextRank	0.456693	5.8	1.37007875
Luhn's Method	0.206278	9.2	0.6188
LSA	0.16816	7.5	0.5044843

Conclusion

From the above obtained values of F1-score we can say that BERT Extractive Summarizer is the best among the above five summarization techniques. The reason for this being the flat architecture of the BERT extractive summarizer. Hence, if we want to rank the summarizers from most effective to least effective, then the order will be BERT,TextRank,LexRank,Luhn's Method and LSA.

References

1. https://chriskhanhtran.github.io/posts/extractive-summarization-with-bert/
2. https://pypi.org/project/lexrank/
3. Parth Mehta, Prasenjit Majumder,Effective aggregation of various summarization techniques,Information Processing & Management,Volume 54, Issue 2,2018,Pages 145-158,ISSN 0306-4573,https://doi.org/10.1016/j.ipm.2017.11.002.(http://www.sciencedirect.com/science/article/pii/S030645731630632X)
4. https://iq.opengenus.org/luhns-heuristic-method-for-text-summarization/