Analysis of NLP Techniques for Text Document Summarization

Kapilesh Yadav Bongu
ECSS
University of Texas at. Dallas
Dallas,U.S.A
KXB210026@utdallas.edu

Shiva Nandini Reddy
ECSS
University of Texas at Dallas
Dallas, U.S.A
SXN210079@utdallas.edu

Aneena Manoj
ECSS
University of Texas at Dallas
Dallas, U.S.A
axm210187@utdallas.edu

Chitteti Rishitha
ECSS
University of Texas at Dallas
Dallas, U.S.A
RXC210003@utdallas.edu

***Abstract*— The practice of extracting the most crucial information from an individual or combined source to create a condensed version for a specific user(s) and the task is known as Document Summarization. There has been a lot of evolution in the techniques for text summarization and the most recent advancements include using Natural Language Processing (NLP) techniques. This report discusses a few such algorithms like Latent Semantic Analysis (LSA) and K-means techniques, and the metrics Rouge and cosine-similarity are used to evaluate the performances of each algorithm on a chosen dataset. This report also provides an analysis of the better-performing algorithms between the two. The dataset we used is from the following source.**

Keywords—NLP, LSA Algorithm-means Clustering, ROUGE

I. INTRODUCTION

With the evolvement of technology, there is huge data generated every day every second. There is a tremendous amount of text available, and it keeps expanding daily. This text material needs to be condensed to brief summaries so the key points are highlighted, and the user can skim more effectively rather than reading the entire article. Consider the example of the world wide web (WWW) which is a huge collection of news, articles, blogs, updates, stories and many others so here there is no particular structure to the data so the best method is to search and navigate to the desired result. There is no possibility to manually summarize the text, so we resort to the automatic methods

A. Reasons for why we need automatic text summarizers

Summary cuts down on reading time. They simplify the selection method during researching documents. The efficiency of indexing is enhanced via automatic summarization. Human summarizers are more subject to bias than automatic summarizing systems. The fact that they offer individualized information makes personalized summaries effective in question-answering systems. Commercial abstract services can process more texts since they can use automatic or semi-automatic summation methods.

B. Approaches to Summarizing text documents

The various aspects of text summarizing can be broadly divided into three categories dependent on their input type single or multiple documents, purpose, general, particular, or query-based and type of output be it extractive or abstractive. There are two primary methods for pruning text documents

1) Extractive Methods:

The process of extractive text summarization entails choosing words and phrases from the original text to include in the extracted summary. Techniques include rating importance of sub subsentences to select only those that are most pertinent to the source's message. The process of extracting text is comparable to that of skimming, in which the summary, headings as well as subheadings, statistics, and the first and final lines of text of a paragraph, are all read before deciding whether to read the entire document in detail[1]. Additional instances of extraction include crucial textual segments related to clinical significance such as patient/problem, intervention, and result.

2) Abstractive Methods:

Abstractive text summarization entails creating wholly original words and sentences that accurately convey the meaning of the source material. Although more difficult, this strategy is finally employed by humans. The way traditional approaches work is by picking out and compressing certain text from the source document.

To condense a text than extraction, abstraction may change the extracted content more significantly by paraphrasing portions of the source document. However, such a process includes both Natural Language Processing (NLP)[4] and frequently a thorough comprehension of the original text's subject matter in cases when the original document relates to a specialized field of knowledge, making it computationally considerably more difficult than extraction.

Because it is simpler, extractive methods have historically produced the most successful text summary techniques; however, abstractive methods provide the possibility of more comprehensive answers to the issue.

C. Background Work

Recent years have seen a significant improvement in the latter challenge as a result of neural sequence to sequence models. The Pointer Generator Network (PGN)[2] with the copy mechanism is now the most used method for producing summarizing text. Our research is comparable to that of [5], who suggests extracting and abstracting the abstractive summarization problem into two subtasks. From the document sentences, the extractor module extracts k sentences. Each of these k statements is taken by the abstractor, who then paraphrases it to resemble a concluding summary sentence. This method has a flaw in that it makes the assumption that each summary phrase is derived from a single sentence in the original source. A summary phrase really span several phrases inside a same page that are all on the same subject. Another problem is that it only takes into account k (often 3-5) sentences out of the total document's 40–50 phrases when creating the summary, which results in

a major loss of information. These problems are lessened by our method.

II. NLP Techniques for Document Summarization

There are several NLP techniques for text summarization but the ones we are using are Latent Semantic analysis (LSA) and K-means.

A. Latent Semantic Analysis(LSA)

LSA is a process for extracting document summarization called latent semantic analysis that uses unsupervised learning. By using singular value decomposition on the term-document frequency matrix, it extracts sentences with semantically important words. SVD, a mathematical method, is used to represent a large piece of the document into a matrix with the count of document words as (rows represent unique words and columns represent each document). This technique reduces the number of rows while maintaining a similar structure among columns[3]. Ay represents the weighted words of sentence k in the document, and A = [A1, A2,..., An] is produced. In a text with m words and n sentences, A is a m x n matrix. The SVD of a m x n matrix A, where m > n, is defined as where V is a n x n orthogonal matrix, U is a m x n column-orthonormal matrix with left-singular vectors, and I is a n x n diagonal matrix with descending non-negative singular values.

The matrix is essentially factorized in SVD. Here, we are keeping the similarity between columns while reducing the number of rows (and thus, the amount of words) (which means paragraphs). A is the document-term matrix in this instance, with documents in the rows (m), singular words in the columns (n), and frequencies where documents and words cross each other. Remember that the original document-term matrix in LSA is approximated by multiplying three additional matrices, namely U,, and VT. The number of elements or subjects in this case is r. The result of running SVD after fixing r (rn) is known as Truncated SVD,[6] and LSA is essentially a truncated SVD alone using equation 1.

$$A = U\Sigma V^T \quad (1)$$

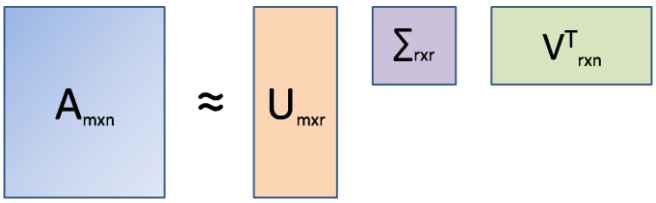

Fig. 1. LSA Algorithm Implementation

B. K-Means

K-means is an unsupervised learning technique that handles the well-known clustering issue. The process classifies a given data set using a specified number of clusters (assuming k clusters). The primary concept is to define k centroids, one for each cluster. These centroid are chosen to be as far apart as feasible. The following step is to associate each point in a given data set with the closest centroid. When all points have been categorized, we compute k new centroid as the new centers of the clusters formed in the previous phase.

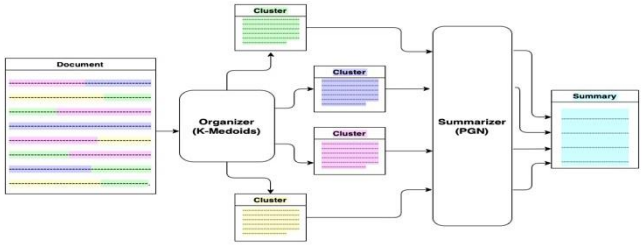

Fig. 2. K-Means Algorithm Implementation

The primary concept is to define k centroids, one for each cluster. These centroids are chosen to be as far apart as feasible. When all points have been categorized, we compute centroid with are of count K as the new centroids of the clusters formed in the previous phase. The different association is created between the same data set points and the closest new centroid after we obtain these k new centroids. Up until there are no more modifications, the k centroid moves in each step. Although the K-means algorithm will always come to an end, it does not always select the configuration that corresponds to the global objective function minimal as the best configuration. The initial, randomly chosen cluster centers have a significant impact on how sensitive the algorithm is.

III. Implementation

The project is implemented using two algorithms namely LSA algorithm and the K-means clustering. Initially the raw data is extracted from the input source, the raw data is pre-processed to remove all the unwanted characters and empty spaces. Later the algorithmic implementations are done from the scratch and the pre-processed data analyzed based on the three different techniques and the number of lines for the output are mentioned. After successfully implementing the code, we get the desired summary. Now that we got the summary of the documents, we evaluated the obtained summary with the reference summary using ROGUE metric and compared the results of the metrics for the two implantations.

A. DataSet:

The dataset we used for the project is We use the non-anon of the CNN / DailyMail dataset [4] modified for summa- rization. It consists of various articles from the CNN and DailyMail websites and their bullet highlights. These highlights are considered as the gold standard summaries. The number of training, validation and test examples are 277, 226, 13, 368 and 11, 490 respectively where the news articles contain an average of 780 tokens (roughly 30 tokens per sentence), and the gold standard summaries contain an average of 56 tokens. The mean of the number of sentences per summary is 3.75. This observation is the motivation for deciding the number of output sentences we want in the summary

B. Latent Semantic Analysis (LSA) Implementation

We followed the LSA algorithm to implement the document summarization. There are 3 steps to achieve the result.

*a) Preprocessing Data:*Initially, the input data is cleaned by removing the stop words using NLTK library and

the data is cleared by deleting the null values, and special characters and assigning the reference summary.

b) Singular Value Decomposition(SVD): The SVD computation is implemented and the truncated SVD matrix is generated.

*c) Summarized Output:*The summary is hereby generated by taking the diagonal matrix and then identifying the important words that form part of the sentence and hence the summary is achieved.

The flowchart of how LSA is implemented is shown fig.3

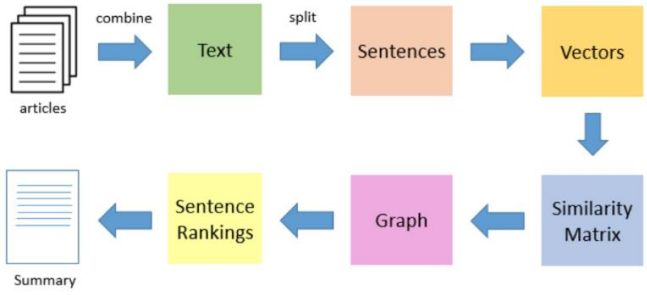

Fig. 3. Flow Chart of LSA Algorithm

C. K-means Clustering Implementation

We followed the K-means Algorithm clustering to achieve the text summarization. There are three steps involved in this implementation.

a) Preprocessing Data: The input data is filtered out from the null values, empty spaces , stop words and other special characters.

b) Word Vector generation: After cleaning the sentences the word vectors are generated.

b) Sentence Vector generation: Now with the obtained word vectors sentence vectors are generated

c)Centroid: The centroid is calculated using K-means clustering. After the creation of sentence vectors, we will model and perform clustering to arrange sentence embeddings into a pre-defined number of clusters equal to the required number of sentences in our summary. In this situation, I set the number of centroids to be equal to the user input. Each cluster of sentence embeddings may be viewed as a set of semantically similar sentences that carry the same information and whose meaning can be expressed by only one phrase from the cluster. The sentence vector with the shortest euclidean distance from the cluster centroid represents the entire group.[9] These sentences from each cluster are organized in the same way as the original text to generate a comprehensible summary.

d)Sentence Extraction: The nearest and best sentence vector is selected from the cluster centroid and hence the summary is results.

D. Evaluation Metric - ROGUE

ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[7] are the accepted evaluation metrics that we employ. These strategies only take into account syntactic overlap. We compute pair-wise ROUGE-L overlap for each sentence in the final summary to assess redundancy, and we use the average recall of the resulting scores to determine the redundancy score for that summary[8]. We compute the ROUGE-L overlap between each sentence in the article and each sentence in the summary to assess coverage.

Another metric used is finding the longest common subsequence (LCS) between a candidate summary and the original reference summary called ROGUE-L. Let us consider X as a reference summary of length m and Y as a candidate summary of length n

Precision is calculated using equation2.

$$P_{lcs} = \frac{LCS(X,Y)}{n} \quad (2)$$

Recall is calculated using equation3

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad (3)$$

F-score is calculated using equation 4.

$$F_{lcs} = \frac{(1+\beta^2)P_{lcs}R_{lcs}}{\beta^2 P_{lcs} + R_{lcs}} \quad (4)$$

Once the summaries are obtained the ROGUE is calculated for reference summary and obtain summary .The ROGUE values are calculated for the two algorithms and the table 1 depicts the results.

IV. Observations and Results

Upon performing the project and evaluations here are the observations and results .

a) Table depicting the Rogue values for the two algorithms.

From the table, we can observe that the rouge values are better for the LSA algorithm and are closer to better results than that of K-means.

TABLE I. ROGUE VALUES FOR K MEANS AND LSA

DATA COLUMN	K MEANS			LSA		
	ROUGE -1	ROUGE -2	ROUGE-l	ROUGE-1	ROUGE-2	ROUGE-l
1	r: 0.1641791044776119 4, p: 0.1428571428571428 5, f : 0.1527777728018906	r: 0.05263157894736842, p: 0.04878048780487805, f: 0.05063290639961593 6	r: 0.1492537313432835 7, p: 0.1298701298701298 6, f: 0.1388888839130017	r: 0.1194029850746268 6, p: 0.3478260869565217 3, f: 0.1777777739728396	r: 0.03947368421052631 4, p: 0.125, f: 0.05999999635200021 4	r: 0.1194029850746268 6, p: 0.3478260869565217 3, f: 0.1777777739728396

2	r: 0.2876 712328 767123 , p: 0.3333 333333 333333 , f: 0.3088 235244 387976	r: 0.09375 , p: 0.11688 311688 311688, f: 0.10404 623783 487613	r: 0.2328 767123 287671 , p: 0.2698 412698 412698 , f: 0.2499 999950 270329 3	r: 0.2876 712328 767123 , p: 0.2763 157894 736842 4, f: 0.2818 791896 328995	r: 0.09375 , p: 0.09090 909090 909091, f: 0.09230 768730 887601	r: 0.2465 753424 657534 , p: 0.2368 421052 631578 8, f: 0.2416 107332 570606 2

b) Reference summary and Obtained Summary

Here are the obtained results of summaries generated by LSA model and K-means and the original reference. By the evaluation metrics we can see that the LSA algorithm performs better than K-means and the summaries are more versatile and unique.

Summary generated by our LSA model:

"Many others, however, said they were in shock and disbelief over the allegations. One man complained he was deep in debt after paying the school a large amount of money to board his son. Captive boys and men were rescued from an Islamic religious school in Pakistan They were reunited with their families this week The facility was a school and drug rehab clinic Authorities say they're searching for the owners; three others arrested at the facility"

Summary generated by K-means model:

"During the day, they worked and did religious studies. But the future of the rescued children was unclear. Captive boys and men were rescued from an Islamic religious school in Pakistan They were reunited with their families this week The facility was a school and drug rehab clinic Authorities say they're searching for the owners; three others were arrested at the facility"

Reference Summary:

The 54 men and 14 boys rescued after being found chained this week at an Islamic religious school in Pakistan have been reunited with their families or placed in shelters, authorities said. The group was discovered in an underground room with heavy chains linking them together. The school, Al-Arabiya Aloom Jamia Masjid Zikirya, which also was a drug rehab clinic, is in Sohrab Goth, a suburb of Gadap in Karachi.

V. TECHNOLOGIES AND PLATFORMS USED

1) File storage: For storing our dataset file, we utilized Amazon's AWS S3 bucket. The data file "s3:/cs6350utd/cnn/" was uploaded to an S3 Bucket.

2) Language and environment of coding: Pyspark was used to write our software. It was run and tested on the Apache Spark environment using Databricks.

3) NLP tools: To tokenize our data, we utilized the NLTK package. We also eliminated stopwords from our list of tokens using the NLTK stopwords set from its corpus.

4) Python tools:

• Numpy, a Python library, was widely used for performing operations on matrices (such as einsum, normalization, and so on).

VI. RELATED WORK

The area of automatic text summarization has received a great deal of attention and has made considerable advancements. For the purpose of handling both extractive and abstractive summarization, many methods have been put forth. Due to its simpler setup, the extractive setting has received more attention in earlier works. In these studies, several strategies utilizing graphs and ILP (integer linear programming) have been developed[5].

Deep learning approaches have largely been responsible for the success with which the job of abstractive summarization has been met. To create summaries, Liuetal present a generative adversarial network-based methodology. Numerous domain-specific publications have advocated summarizing as a method of analysis, including temporal summarization, movie review summarization, customer review summarization, and book summary[10]. Furthermore, both extractive and abstractive settings have been addressed using reinforcement learning. On text summary, there is a wealth of literature, including several studies that give a thorough review of the issue.

VII. CONCLUSION AND FUTURE WORK

In this study, we have implemented two extractive summarization techniques with k-means where the sentences are organized into three clusters and one sentence from each is picked. And the Latent semantic analysis divides up significant text documents into n parts, each of which provides a distinct perspective of the message conveyed by the text. Then the top keywords are selected to construct a summary. Both models are evaluated using the rouge metric and the LSA is found to better in most of the summarizations.

The K-means algorithm can be used to identify and group the sentences that infer similar ideas into one cluster later abstractive summarization techniques like PGN can be applied to generate a better summary.

This project can be further extended by analysing more NLP algorithms like Luhn Algorithm,KL-sum and many others.Also Machine learning techniques can be coupled like neural networks,Recurrent Neural Networks(RNN) and so on. Further, various evaluation matrices like cosine similarity, perplexity and metor can be used for better analysis apart from ROGUE.Another fascinating future work would be our suggestion on testing our strategy with several summarizers to see if the benefits in redundancy and coverage are the same across summarizers.

VIII. References

[1] Regina Barzilay and Michael Elhadad. “Using Lexical Chains for Text Summarization”. In: *Intelligent Scalable Text Summarization*. 1997. URL: https://aclanthology.org/W97-0703.

[2] Alexander R. Fabbri et al. *Multi-News: a Large- Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model*. 2019. DOI: 10.48550/ARXIV.1906.01749. URL: https://arxiv.org/abs/1906.01749.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Nat- ural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

[4] https://www.tensorflow.org/datasets/catalog/cnn_dailymail

[5] Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Re- cent advances in document summarization. *Knowl- edge and Information Systems*, 53(2):297–336.

[6] Chin-Yew Lin. 2004. ROUGE: A package for auto- matic evaluation of summaries. In *Text Summariza- tion Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[7] Rada Mihalcea and Hakan Ceylan. 2007. Explorations in automatic book summarization. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pages 380–389.

[8] Sarita Mandal et al. “Extractive Text Summa- rization Using Supervised Learning and Natu- ral Language Processing”. In: *2021 International Conference on Intelligent Technologies (CONIT)*. 2021, pp. 1–7. DOI: 10.1109/CONIT51480.2021. 9498322.

[9] Elena Lloret and Manuel Sanz. “Finding the Best Approach for Multi-lingual Text Summarisation: A Comparative Analysis.” In: Jan. 2011, pp. 194– 201.

[10] N. Nazari and M. A. Mahdavi. “A survey on Au- tomatic Text Summarization”. In: *Journal of AI and Data Mining* 7.1 (2019), pp. 121–135. ISSN: 2322-5211. DOI: 10.22044/jadm.2018.6139.1726. eprint: http://jad.shahroodut.ac.ir/article 1189 28715967fcd8b7bfb463ab90aca5a9f7.pdf. URL: http://jad.shahroodut.ac.ir/article 1189.html.

IX. Appendix

All 14 boys were returned to their families, senior police official Ahsanullah Marwat told CNN. Of the adults, 47 had been released to their families, and seven were handed over to a shelter for the homeless, he said. Three people who worked at the facility were arrested, but the four men who ran the place were still at large, Marwat said. Officials said the facility was part madrassa and part drug-rehab facility, and the captives were chained at night apparently to prevent their escape. "The operation was successful, and we plan on continuing our work to ensure that places like this are shut down," Marwat said. Many of the captives told police their families sent them there because they were recovering drug addicts. During the day, they worked and did religious studies. But the future of the rescued children was unclear. One woman told a local television station that she was willing to pay the police to keep her troublesome child. She said she would rather have the facility remain open, regardless of how it treated the children. Many others, however, said they were in shock and disbelief over the allegations. One man complained he was deep in debt after paying the school a large amount of money to board his son. Captive boys and men were rescued from an Islamic religious school in Pakistan They were reunited with their families this week The facility was a school and drug rehab clinic Authorities say they\'re searching for the owners; three others arrested at the facility