Obstacles in Automatic Text Summarization using Natural language Processing

Jeevana Pathipati

University of Cincinnati, pathipja@mail.uc.edu

ABSTRACT:

Automatic text summary is thought to be the prior challenge in Natural Language Processing (NLP). For example, the amount of text on the internet is expanding exponentially. Now, we're talking about a difficulty or a challenge with automatic text summarization. In this process, there are two forms of output. Extractive text summarization and abstractive summarization are the first and second techniques. Insufficient coherence and consistency are issues in extractive summarization, and additionally there are complications with abstract summarization. Entity Hallucination, irregularity, and non-factual as we understand the definition of Illusion as assuming something to happen that hasn't actually occurred. As database sources, we employed the "IEEE explorer", and the "ACM" digital library in order to finish this study and acquire a well-regulated outcome. In the "query box" of each of the three databases, we entered our first query using the "Automatic text summarization" option. Following a review and examination of the findings, we proceeded with the challenges that were associated with it. In addition, because of the criteria, we have opted to include research articles that largely focus on the many forms of text processing and the challenges they encounter, as well as what the conclusions of this study are. To Complete this paper with accurate results author applied systematic literature review methodology, Over the course of the past 10 years, we have extracted all of the research articles for use as references from each of the two databases. An effective text summary system should be able to grasp the full text, reorganize the information, and provide summaries that are coherent, informative, and amazingly concise in order to communicate the main ideas contained within the original text. This issue occurs due to the fact that the system has developed sufficient intelligence to recognize and retain all of the data that is now available in it. The deep learning algorithm has made significant progress in text summarization, but it is not without flaws, including the occurrence of hallucinations. This occurs as a result of the cutting-edge deep neural network theory that is used in deep learning. This model gives the machine or systems with amazing characteristics that it recognizes and gathers the data from the information that came before it.

Additional Keywords and Phrases: Extractive (1), Abstractive (2), Hallucination (3), Deep learning (4), Proposed models (5), illusions (6), deep learning (7)

1 INTRODUCTION

The process of summarizing a text is selecting the most relevant parts of a piece of writing and putting them into a short summary. In recent years, there has been a greater need for summarizing. In news, business, and research, it is very important to have a short, clear summary of the information. The challenge of autonomously summarizing text is well-known in the Natural Language Processing (NLP) area. Extraction summarization and abstraction summarization are the two primary categories of text summarizing methods [8]. Extractive summarization is a way to make a summary without reworking or rephrasing the actual document. It works by finding the most important sentences or phrases in the source text and grouping them together. This kind of summary, known as abstractive summarization, requires the writer to have a more in-depth familiarity with the original material and to craft their own original phrases to best convey the essential ideas presented therein. This makes the summary better by limiting redundant information and concentrating on the meaning of the source text. Summarization—identifying and emphasizing the most significant information from several sources—is increasingly commonplace. Newscasts keep many individuals updated about global events. Stock market reports help them invest.

Online ratings and reviews impact their movie choices. Skimming summaries before choosing may save time. As online content expands, summarizing technologies are making it harder to provide meaningful and timely summaries. Academics study content extraction and abstracting methods. The authors divide data synthesis methods into two categories. Knowledge-poor approaches must not change rules for each new usage or language [5]. Knowledge-rich techniques assume that a deeper understanding leads to a better summary. They need to learn, update, and adapt a large set of rules to new settings and languages. The authors think summary generators will help conquer massive information universes. Many research approaches have been suggested to enhance seq2seq models, allowing them to deal with difficulties such as efficiency, human comprehension, and intensity, and produce summaries of a very great quality. The majority of these approaches deviate in anyone of the three ways: variable inference, network structure and decryption." Neural Abstractive Text Summarizer" toolkit was created as a free and open-source framework for abstractive text summary. Issues like training a model quickly and effectively in parallel are also important. In this paper, we give an assessment of the literature on sequence-2 sequence designs for abstract text summing up from the perspectives of deep learning architectures, training methods, and summary production algorithms [1]. The field of information systems has made significant progress in the area of automatic summarization of text. In order to provide the logical and informative overview, most of the techniques need on expert-level topic knowledge. Extractive text summarization involves giving sentences ratings based on certain characteristics. Researchers have presented a plethora of feature-based scoring approaches for autonomous text extraction and summarization. In this study, we look at the factors that go into sentence grades. In this article, we explore the findings on the use of feature combinations for scoring [2].

The resulting summary is compared against an abstract summary of the DUC 2002 dataset using ROUGE-N. Despite the fact that we are evaluating our extracted summary to the Gold Overviews, which are abstracted versions of the DUC dataset., we found that Comb1 was ROUGE-1's best feature set, with an F-Measure of roughly 70%. Word frequency, upper case letters, proper nouns, and word co-occurrence are some of the techniques that are used in extractive text summarization. Similarity in lexicon, "cue-phrase," and Inclusion of numerical data, Length of the sentence, location inside the sentence, and prominence of the sentence Similarity to the name of the book, Aggregate similarity, Score for the Text Rank, Path hidden by bushes [3]. The exponential development of the internet has led to a substantial rise in the quantity of information that can be accessed, particularly in terms of text documents. It is no longer possible to effectively sift through the vast amount of materials that are available on the Internet and in digital libraries in order to find information that is valuable. The difficulties associated with poor levels of efficiency and accuracy while working with lengthy text include the following: their capacity is insufficient to cope with extremely long input; they are unable to precisely replicate factual facts; and they have a tendency to repeat themselves. In this study, we present a hybrid model that incorporates both extractive and abstractive approaches [4].

Even during extraction stage, we can build a graph and recommend a hybrid statement measure of relevance that combines phrase vector and Levenshtein.This is done in the "extractive" phase. Then, after using this measure to rank the sentences, Choose the most important sentences and string them together to make a clear indication that will be fed into the summarized system. A traditional seq-to-seq selective attention structure is changed in two ways in the section about abstraction [9]. CNN/DM is a corpus of single-text summaries with several sentences. When it comes to abstract text summary, CaPE, or Contrastive Parameter Ensembling, is used to lessen hallucination. To decrease the risk of hallucinations, the new CaPE training approach makes strategic use of the noise changes in the sample. NATS [10] Neural Abstractive Text Summarization with seq to seq is a tool that does framework, varying inference, decoding or generation. Several models have been suggested for modeling languages and doing things like language processing,but they all have their own set of issues, such as the repetition of words and the lack of consistency. Extractive summaries are often generated using automated text summarizing using sentence-based extractive summarization algorithms. On determine which sentences are most crucial when summarizing a material, the conventional approach of summary looks to linguistic characteristics of the sentences themselves. Unlike other approaches to calculating the similarity of sentences, its restriction is unrelated to the semantics of the sentences themselves [6]. Extractive summaries are generated by determining the number of clusters, clustering the sentences in the text using the K-means approach, and then extracting the subject sentences. Each article in the train set is summarized naturally, and two methods for text categorization are presented based on these summaries in order to minimize the size of feature vector space and the compute difficulty of classification [7]. The summation is utilized in place of the original text for selecting and categorizing features; in the second method, the summaries are used to pick and weight features for individual documents, and the KNN algorithm is used to classify unstructured texts.

Research Question: It is getting more difficult to produce accurate and timely summaries when there is a greater volume of internet information. How may these difficulties with text summarization be resolved?

2 METHODOLOGY

The methodology used in this study is called a Systematic Literature Review, and it focuses specifically on discussing the problems with Automatic Text Summarization and the consequences of those problems. In order to finish this research and get useful findings out of it, we made use of the ACM (Association for Computing Machinery) and IEEE Xplore databases as our primary sources of information. In January of 2023, we started looking for research publications that were connected to our project. We conducted a search using a query in order to find appropriate research articles, and we also used a query to search our search history for publications that were related to our topic. Some of the keywords that we used were "text summarizing," "natural language processing," "Abstractive text summarization," and "Extractive Summarization." These libraries were selected by us because of their status as standard libraries whose publications have been vetted and approved by relevant authorities before being made available to the public.

In order to answer the study topic, we conducted the thematic analysis of the data using six different publications. The summaries of these 6 articles are the primary focus of the study of the data, and qualitative research techniques are used to discover common themes. The results were derived through a thematic evaluation of the data, which revealed commonalities among the six articles chosen for this study.
The fact that these libraries are equipped with a brilliant module that is known as advanced search is the primary advantage that they provide. Within advanced search, users have the ability to conduct searches with greater specificity by collecting the references in such categories as publication title, author and abstract. Additionally, the AND keyword is applied to connect the queries and give users the results that are most related to them and can even filter the dates as per our requirement.

You will find a great number of papers that correspond to your search; however, we have decided to use this particular paper as our reference. This is due to the fact that it is extremely similar to the research paper that we have been working on, as well as the fact that its abstract was understandable and it produced satisfactory results.

2.1 INCLUSION AND EXCLUSION CRITERIA

The inclusion process for the study, one of the conditions for inclusion was that we included all of the papers that are significant in the study, have been published within the past ten years, and encapsulate all of the necessary as well as additional keywords, such as "Automatic Text Summarization", "Abstractive", "Extractive", "NLP". For exclusion criteria, all of the publications that did not include important keywords like "Deep Learning", "Models", "Automatic Text" have been omitted from consideration since this criterion is one of the methodological limitations. discarded the documents since they were not pertinent to the discussion at hand. In addition, we did not consider any findings that had been published later than ten years. Only those written in English were taken into consideration.

2.2 SEARCH PROCESS

Utilized the ACM and IEEE Xplore libraries as our main data sources and did a search using a query to locate acceptable research papers; we also used a query to browse our search results for papers that were applicable to our subject.

2.3 QUALITY ASSESSMENT

Six distinct sources were used to complete the thematic analysis of the information. Quantitative research methods are utilized to find recurring themes in the abstracts of these 6 papers.

2.4 DATA EXTRACTION

The information was gathered using a thematic analysis of the data, which showed similarities across the six articles used in this investigation

The following ACM libraries were utilized for this paper:

[Abstract: automatic] AND [Abstract: models] AND [All: text summarization] AND [Title: neural] AND [E-Publication Date: (08/01/2017 TO 03/31/2023)] When we applied this syntax, the results showed up with 279 publications, and after applying other filters, it was narrowed down to double digits; among them, we picked one article to use as a reference, and that paper is entitled as Ref-1.

[Abstract: automatic] AND [Abstract: text summarization] AND [Abstract: methods] AND [Title: analysis] AND [E-Publication Date: (02/01/2010 TO 01/31/2022)] after using this syntax, the findings show that there are 87 articles; these results were obtained by reading the content of the research papers that were selected to serve as our reference, which is Ref-2.

[Abstract: summarization] AND [All: natural language processing] AND [Abstract: structured text] AND [Abstract: information] AND [E-Publication Date: (10/01/2014 TO 07/31/2022)] and the outcomes of applying this syntax 346 results have been produced, and among them we have filtered and selected one as the source for our study, it is Ref-3.

[All: challenges] AND [Abstract: facing] AND [All: automatic text summarization] AND [Title: mechanism] AND [E-Publication Date: (09/01/2015 TO 09/30/2022)] The results of applying this syntax are now known, and among them, we have selected one to serve as a reference for our work; its number is Ref-4.

[Title: challenges] AND [Title: automatic summarization] AND [E-Publication Date: (08/01/2010 TO 09/30/2022)] this syntax was utilized, and after filtering using the title, it results the output included 23 publications; among those articles, we chose one to serve as our reference, and that document is referred to as Ref-5.

[Title: natural language processing] AND [Title: ai] AND [All: abstract:] AND [E-Publication Date: (08/01/2012 TO 05/31/2022)] and the findings were achieved by reviewing the text of the research papers that were chosen to serve as our reference, which is referred to as Ref-6. The data demonstrate that there are a total of 32 publications after using this syntax.

[Title: trainable] AND [Title: summarizer] AND [E-Publication Date: (07/01/1995 TO 03/31/2022)] and we applied this syntax in ACM in order to attain better results, and we chose one of the publications as our reference; that is referred to as Ref-9.

[Title: automatic text summarization] AND [Title: survey] AND [E-Publication Date: (09/01/2011 TO 03/31/2022)] Since we knew this information from the reference that was found in reference paper 1, and because we are familiar with the title, we utilized it in this instance as a keyword in advanced search and referred it as Ref-10.

This article makes use of the following IEEE Xplore library:

("Document Title":Clustering) AND ("Document Title":extraction) AND ("Abstract":automatic text summarization) Filters Applied: 2010-2022 and we used this syntax in an advanced search on IEEE Xplore, and it produced 138 results; from those 138 results, we selected one document as our reference that paper is referred to as Ref-7.

("Abstract":KNN Algorithm) AND ("Abstract":text categorization) AND ("Abstract":summarization)) Applied Filters: 2010-2022 and we utilized this syntax in IEEE Xplore in order to get efficient results, and we choose the publications, which is referred to as Ref-8.

3 RESULT

Taking into account all of the research papers, each one of them created or presented a systematic literature review approach in order to generate the qualitative analysis in an effective summary. The outcomes of the one approach that was described in our reference are shown in this part; it is possible that this Systematic Literature methodology is the most effective one based on the qualitative analysis of the study. Furthermore, we are providing a comparison analysis in the form of a table, with one table including the suggested methodology and the other table holding the existing methodology, that has not been changed, and we are presenting the core characteristics together with the related functions. The second-best technique in order to achieve a concise summary is by using the CaPE model. We give its findings a tabulated format, and the values will be defined, and then we compare them with the base and CaPE model. According to the findings of the research, the CaPE model achieves a 15–16 percentage improvement in efficiency.

$$\text{Precision} = \frac{\text{(No of words that appear more than once)}}{\text{(Total no of words in the summary)}} \quad \text{, Where P = Precision}$$

$$\text{Recall} = \frac{\text{(No of words that overlap)}}{\text{(Total no of references used)}} \quad \text{, where R= Recall}$$

The term "recall value" refers to a value that is equivalent to the total number of relevant keywords that are produced in (summary) result.

$$\text{F measure (F)} = \frac{2(PR)}{(P+R)}$$

The F measure, often known as the F value, is derived from the addition of the P and R quantities; to put it another way, we may define F as the overall sum of the P and R measurements added together.

Table 1: The values of Rouge-1 and Rouge-N evaluated by NATS Model.

Number of Documents in scale of 10		10.00	5.00	3.00
Rouge-1	Recall (R)	43.00	45.00	46.00
	Precision (P)	35.00	35.00	36.00
	F-measure (F)	58.00	59.00	58.00
Rouge- N	Recall (R)	36.00	37.00	38.00
	Precision (P)	30.00	31.00	32.00
	F-measure (F)	46.00	47.00	48.00

Where Recall R, Precision P, F-measure F are measured in percentage (%).

Table 2: The values of Rouge-1 and Rouge-N evaluated K-means Clustering Model.

Number of Documents in scale of 10		10.00	5.00	3.00
Rouge-1	Recall in %	37.00	37.00	38.00
	Precision in %	30.00	29.00	30.00
	F-measure in %	51.00	50.00	52.00
Rouge- N	Recall in %	28.00	27.00	28.00
	Precision in %	23.00	23.00	23.00
	F-measure in %	36.00	35.00	36.00

D Arc describes the amount of dependent data in the output text that is assumed to be associated with the source article. Dsum determines how often summaries are free of dependency arc problems,.E-Psrc is a metric that measures how many entities in the abstract can be found in the primary source. E-Rref calculates the coverage rate, the percentage of cited entities that are also present in the summarized version.The BERTScore's (Zhang et al., 2019) recall (accuracy) is denoted by BS-P (R). QEval is a quality-assurance-based measurement of the reliability of claims (Scialom et al., 2021). To get the entailment score, MNLI use the Robert large model that was trained using the MNLI dataset (Williams et al., 2018). QAFactEval is a supplementary QA-based factual consistency indicator that improves the question filtering and response overlap functions.

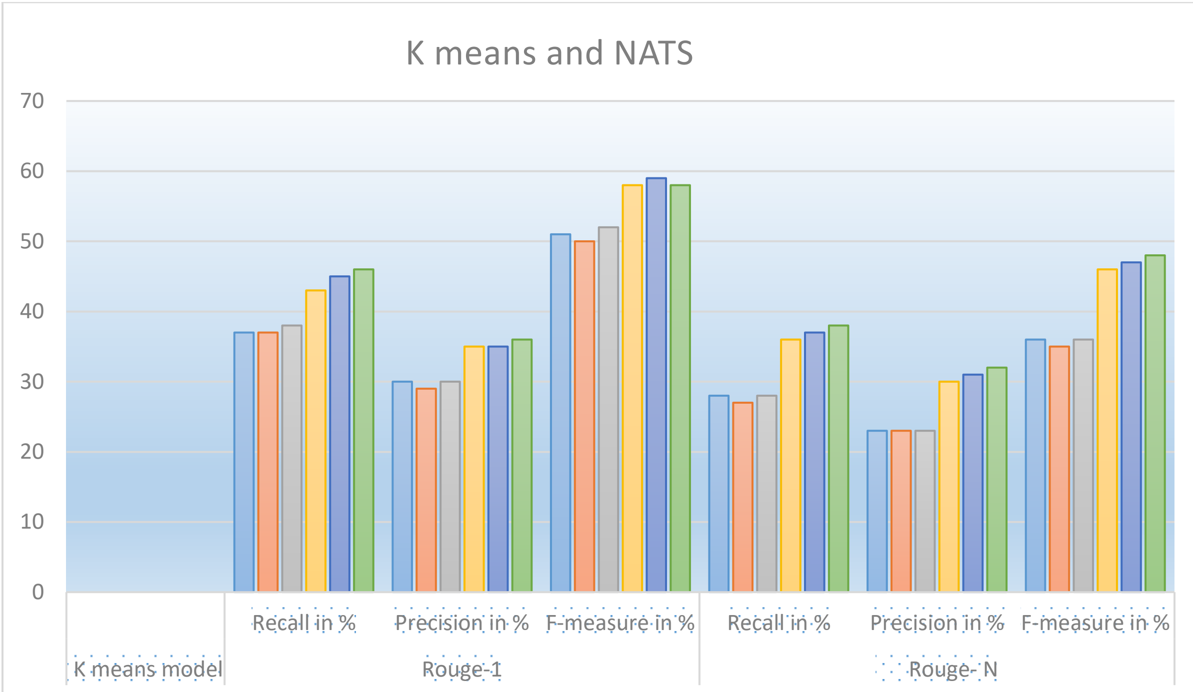

Figure 1 represents gives the Comparision outcomes of K-means clustering and NATS Models

Table 3: CaPE Model vs. Base Model Comparison

Base Model	E-P ref	9.51
	E-R ref	6.74
	Darc	9.77
	Dsum	6.19
	QEval	5.95
	R1	9.30
	R2	8.29
	BS-P	4.42
	BS-R	2.12
CaPE	E-P ref	9.82
	E-R ref	8.68
	Darc	9.88
	Dsum	5.83
	QEval	6.01
	R1	9.37
	R2	8.28
	BS-P	4.37
	BS-R	2.08

Table 4: Comparison of the Base Model and the CaPE Model using FactEval Metrics, MNLI, CNN/DM

CNN/DM	MNLI	QA FactEval
Base Model	8.42	4.55
CaPE Model	8.68	4.60

4 DISCUSSION

In this paper, we can say that there are many approaches that follow different algorithms and use different ways of relationships to get better results. As we said in the introduction, there were a lot of problems with Automatic Text Summarization, and in each reference, the authors came up with a new way to summarize in a way that is factual, consistent, relatable, recognizable, and effective. In this paper, we look at the problems with summarizing text. Since our university gave us permission to use digital libraries, we asked for their help to finish this research paper. [8] In this paper, we have to figure out what the problems are and what the best way is to solve them. As we all know, the amount of data on the Internet is growing every day. Every millisecond, a huge amount of new data is added.

The limitations faced by the author to get same proposed model analysis in different research paper's and author mentioned different different proposed models like cape, knn, k measure, sequence to sequence model all these models were calculated based on different categories which becomes hard to make comparision and other limitation found by author is According to the research papers we chose as sources for our paper, each author first identifies the problem and records all of its experimental values. Then, they suggest a way to fix the problem. For example, in the paper we talked about above, To address the issue of unusual vocabulary, the author proposed a pointer technique using seq-to-seq. A number of issues remain, nonetheless, with this concept.

For abstract text summarization, we evaluate the literature on sequence-2 sequence models from the point of view of systems, training methods, and algorithms for making summaries.[1] Extractive summaries are often generated using sentence-based extractive summarizing algorithms in automated text summarization systems. The traditional method of summarization uses linguistic properties of sentences to identify the most important ones for a given piece of content. Its constraint has nothing to do with the meaning of the phrases, unlike other methods of determining how similar they are.[5]

In this case, the data is limited in a way that is qualitative analysis. This is because we chose 6 articles from the papers that came out of the search using a random sampling method. These articles were research papers from an electronic database, and the abstracts of these papers told us that they were being used all the time.

Author suggests the K measure clustering for extractive text summarization as the model reduces the humiliation and get accurate summarization with meaningful sentence without avoiding or changing the content information.[6]

Tabular data is the focus of this paper. We can observe that the parameter of F-measure is greater than the author's suggested model than for the current model by comparing Table 1 (results for the author's proposed model) and Table 2 (results for the existing model). What this signifies is that the author's suggested model is effective, as shown by the tests. [3]
And finally, we might claim that here is where all the issues with text summarization, both in abstractive and extractive forms. We also looked at the ideas of the authors whose papers we used as references and made sure we understood their results and how they got there. With the help of libraries, we were able to finish our paper well.

REFERENCES

[1] Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2021. Neural Abstractive Text Summarization with Sequence-to-Sequence Models. ACM/IMS Trans. Data Sci. 2, 1, Article 1 (February 2021), 37 pages. https://doi-org.uc.idm.oclc.org/10.1145/3419106

[2] Yogesh Kumar Meena and Dinesh Gopalani. 2014. Analysis of Sentence Scoring Methods for Extractive Automatic Text Summarization. In Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies (ICTCS '14). Association for Computing Machinery, New York, NY, USA, Article 53, 1–6. https://doi-org.uc.idm.oclc.org/10.1145/2677855.2677908

[3] Steven J. Simske and Rafael Lins. 2018. Automatic Text Summarization and Classification. In Proceedings of the ACM Symposium on Document Engineering 2018 (DocEng '18). Association for Computing Machinery, New York, NY, USA, Article 1, 1–2. https://doi-org.uc.idm.oclc.org/10.1145/3209280.3232791

[4] Yong Zhang, Erdan Chen, and Weidong Xiao. 2018. Extractive-abstractive summarization with pointer and coverage mechanism. In Proceedings of the 1st International Conference on Big Data Technologies (ICBDT '18). Association for Computing Machinery, New York, NY, USA, 69–74. https://doi-org.uc.idm.oclc.org/10.1145/3226116.3226126

[5] Udo Hahn and Inderjeet Mani. 2000. The Challenges of Automatic Summarization. Computer 33, 11 (November 2000), 29–36. https://doi-org.uc.idm.oclc.org/10.1109/2.881692

[6] Xiaolian Li and Bo Zhang. 2020. Discussion on Natural Language Processing and AI Poetry. In Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare (CAIH2020). Association for Computing Machinery, New York, NY, USA, 10–13. https://doi-org.uc.idm.oclc.org/10.1145/3433996.3433999

[7] P. -y. Zhang and C. -h. Li, "Automatic text summarization based on sentences clustering and extraction," 2009 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, China, 2009, pp. 167-170, doi: 10.1109/ICCSIT.2009.5234971.

[8] X. -y. Jiang, X. -Z. Fan, Z. -F. Wang and K. -L. Jia, "Improving the Performance of Text Categorization Using Automatic Summarization," 2009 International Conference on Computer Modeling and Simulation, Macau, China, 2009, pp. 347-351, doi: 10.1109/ICCMS.2009.29.

[9] Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '95). Association for Computing Machinery, New York, NY, USA, 68–73. https://doi.org/10.1145/215206.215333

[10] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. Artif. Intell. Rev. 47, 1 (January 2017), 1–66. https://doi-org.uc.idm.oclc.org/10.1007/s10462-016-9475-9