International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 01 | Jan 2022 www.irjet.net p-ISSN: 2395-0072

A Comparative Study of Automatic Text Summarization Methodologies

Shubham Jadhav[1], Ruturaj Mane[2], Rutuja Chaudhari[3], Vishal Chatre[4], Asmita Manna[5]

[1,2,3,4] *Student, Department of Computer Engineering Pimpri Chinchwad College of Engineering-PCCOE Pimpri Chinchwad, Pune-411044, India*
[5]*Professor, Department of Computer Engineering Pimpri Chinchwad College of Engineering-PCCOE Pimpri Chinchwad, Pune-411044, India*

Abstract - *Due to the abundance of available documents and materials all over the internet and various archives, it is becoming difficult for human users to obtain crisp information about a specific topic on time. Therefore, users need to obtain a summary of all available documents instead of a pile of documents. As manual summarization is time-consuming and not cost-effective, automatic text summarization is the need of the hour. Much research has been conducted for automatic text summarization and many solutions have been proposed. However, the automatically generated summaries are still far behind than summaries generated by human users. Most of the research has focused on generating a summary from a single document whereas generating a summary from multiple documents is becoming more important nowadays. In this paper, a survey of existing solutions for automatic text summarization has been presented and the research gaps are identified too. It is hoped that the identified research gaps would help future researchers to channelize the research in the right direction.*

Key Words: **Automatic Text Summarization, Natural Language Processing, Extractive Text Summarization, Language Summarization, Document and Text Processing**

1.INTRODUCTION

Nowadays, it is difficult to find crisp and meaningful information about a specific topic of interest due to the abundance of related and unrelated information all over the internet. Even if the search results retrieve meaningful documents, users may have time-constraint to go through the complete document to understand the gist of the same. Sometimes information is repetitive across multiple documents. Thus, summarizing the major points from multiple documents and crisply presenting that information to the users, is the need of the hour now. However, manually summarizing these huge amounts of information is not feasible or practicable. The Automatic Text Summarization (ATS) is the key solution to this problem.

The ATS aims to produce the summary from single or multiple documents such that the main essence of the document is not missed, and the repetitions are not there. Documents generated by the ATS may not be lossless always, but the major points should not be missed in the abridged summary. The abridged version will help the users to get the key information of the document without much time consumption. Text summarization can be used in many applications for example to generate the meeting summary, news summarization, headline generation, etc.

Many researchers have channelized their energy towards ATS and proposed different solutions but still, there is room for proposing better solutions to satisfy the needs of human users. For example, not many satisfactory works are there to generate summaries from live sports commentary for sports-lovers. There is hardly any method available for generating customizable summaries concerning time constraints as per the sports-lovers' expectations.

In this paper, a comprehensive survey on available approaches and methods of ATS is presented to help the researchers understand the drawbacks of the existing systems and find future research directions. The survey presents different aspects of ATS like approaches, methods, datasets, techniques, and evaluation criteria.

2. CLASSIFICATIONS OF ATS:

Automatic text Summarization is used in various applications and it has two major approaches namely extractive summarization and abstractive summarization. In extractive summarization, the important sentences in the input text document are identified and directly included in the summary whereas in the abstractive text summarization, new sentences are generated from the original text based upon the meaning of sentences. There are three main steps for text summarization: a) Topic identification comprising of word frequency, cue phrases, etc. b) Interpretation and c) Summary Generation. The general steps of ATS architecture are discussed here.

a. Pre-processing: The structured representation of original text is produced using techniques such as tokenization, POS tagging, stemming, etc.

b. Processing: Converting a document into its summary by using different text summarization approaches.

c. Post-processing - This step is about solving the problems in the generated summary sentences.

3. PARAMETERS CHOSEN FOR CLASSIFICATION AND COMPARISON:

In this section, few parameters are identified and defined, based on which research works in the area of text summarization can be compared.

3.1 Classification based on input document size

The text summarization can be performed either from a single document or from multiple documents. Multi-document text summarization uses more than one text input that focuses on a common topic around which a summary is to be generated whereas single document summarization is done from a single document.

3.2 Classification based on the domain:

A domain is the property of the model used to generate the summary. It is classified into domain-specific and general. Some models are developed to create automatic text summarization in a specific domain and that model is not able to produce promising results in other domains. These models are included in the domain-specific models. Unlike domain-specific models, some models are adaptive and are successful in generating summary documents of more than one domain. These are included in the General domain category.

3.3 Classification based on approach

Automatic Text Summarization is broadly classified into two types: Extractive text summarization and abstractive text summarization. In the extractive text summarization, the summary is generated by extracting words or sentences from the input text. For example, an approach can use neural networks to find important words or sentences to be included in the summary to increase their weightage and a general extractive text summarization approach can be implemented to extract those words or sentences to include in the summary. On the other hand abstractive text summarization focuses on identifying the meaning of one sentence and summarizing accordingly.

3.4. Classification based on the result

ROUGE is probably the most important metric for the evaluation of the summaries automatically. ROUGE is an acronym for Recall Oriented Understudy for Gisting Evaluation. The quality of a summary is determined automatically by comparing one summary to another standard summary that is made by humans with the help of various automatic evaluations of the summary. It is based on a metric called BLEU that is defined originally from machine translation which could be applied to evaluate summaries.

3.4.1 ROUGE-N

First, we find out 'n-grams' in the given input. Then we compare the total number of 'n-grams' that are matching within the model-generated text which gives us ROUGE-N. 'n-grams' is nothing but just a collection of tokens or words. When there is a single word then it is called unigram. When there are two words then it is called bigram. When there are three words then it is called trigram. ROUGE-N is based on BLEU which is used in machine translation. BLEU stands for Bilingual evaluation understudy.

A reference is a human-generated summary. ROUGE-N is the overlap that consists of 'n-grams' within the given system and the summaries that are given for reference. ROUGE-1 will refer towards the unigram that has been overlapped within the summary for reference and system given. ROUGE-2 will refer towards the bigrams that have been overlapped within the summary for reference and system given. ROUGE-3 will refer towards the trigrams that have been overlapped within the summary for reference and system given respectively. After it is decided which N to be used then the ROUGE Recall, precision, or F1 score will be calculated respectively.

3.4.2 Recall

When the 'n-grams' are overlapped in the outputs of the model and the reference. Then such an overlapping of total numbers or 'n-grams' is called Recall. After this, a particular number is divided from the overall number of 'n-grams' that are given in the reference.

$$\frac{number\ of\ n-grams\ found\ in\ model\ and\ reference}{number\ of\ n-grams\ in\ reference}$$

$$\frac{count_{match}(gram_n)}{count(gram_n)}$$

The above statements ensure that this model has been collecting the entire information consisting within the given reference. But in contrast, this is not so good at assuring that the given design is failing to push a large number of words outside to gain the score of recall respectively.

3.4.3 Precision

To avoid the drawback of recall the precision metric is used. Precision is basically a performance metric that is applied to the data received by a collection of words. The calculation of Precision is done by dividing the number of 'n-grams' found inside both by the number of 'n-grams' in the model.

$$\frac{number\ of\ n-grams\ found\ in\ model\ and\ reference}{number\ of\ n-gram\ in\ model}$$

3.4.4 F-1 Score

After the above process, the recall values along with the precision values will be received, then they are used to calculate the ROUGE-F1 score.

$$2 \times \frac{precision \times recall}{precision + recall}$$

3.4.5 ROUGE-L

ROUGE-L measures the LCS which stands for Longest Common Subsequence based on the statistics of the output given by the model with the reference given.

3.4.6 ROUGE S

It is a skip-gram concurrency which is a group of some words in its sentence order.

4. APPLICATIONS OF MODELS PROPOSED IN PAPERS

Researchers have worked to summarize texts specific to some particular domains like News, Books, Mails, Medical documents for summarization, etc. Text summarization in a particular domain has some specific appraoches too. In this survey, research papers are also compared based on their application domain.

- News Article Summarization: In case of news articule summarization, some datasets like DUC-2001, CNN, Daily Mail, etc are used. This datasets consist of a certain number of words and some sets of documents. Wordnet is used as it contains English language words. Sets are formed by combining some groups of English words which are known as synsets. It is later parsed with any official news websites like BBC, TOI, etc., to get the information that which a particular sentence should get more importance. Thus, this summary would contain all the important information that is related to the topic.

- Books Summarization: A summary is created to give the content of the book in short, which might save the time of customers to select their book with the genre they want. As the availability of books online increases, more tools are needed to summarize.

- Legal articles Summarization: This can be very useful in legal operations to find a particular rule in very little time and also it would summarize the whole document and the reader would only have to readjust the important part. When a certain legal problem is occurred, lawyers can go through past documents in a very short time through summarization than rather actually reading all the files

- Sports Summarization: In this, the text would be considered from the datasets like SPORTSSUM, BBC sports Article datasets, CNN, etc. First, the articles are selected and then it's performed so that there would be more and more accuracy by avoiding the unneeded part. This would reduce the effort to read the whole article and create just the summary. There could also be the prediction of scores using summary.

Thus, the research papers studied proposed solutions that had a wide range of applications which ranged from small articles like news to large data like books.

5. LITERATURE REVIEW

Automatic text summarization has large number of applications and can be achieved using various methodologies. These methods can produce results but there are chances that they adhere to certain situations. So, literature review helps in accomplishing the task of recognizing the methodology and its best application. The literature survey completed during this survey paper is given below.

There was a paper that proposed the firefly algorithm **[1]** which is a type of swarm-based algorithm. In it, vectors of size N are created and sentences that are to be used in summary are labeled as 1 and remaining as 0. TRF (Topic relation factor), CF (cohesion factor), and RF (readability factor) are used as fitting functions. When we are doing a comparative assessment of extractive summarization **[2]** i.e., TextRank, TF-IDF, and LDA then from the results it is observed that textRank performed better than TF-IDF and LDA. Speech-to-text summarization **[3]** can be performed using extractive text summarization algorithms. First 6 text summarization algorithms: Luhn, TextRank, LexRank, LSA, SumBasic, and KLSum are selected. Then, using 2 datasets, DUC2001 and OWIDSum, with six ROUGE metrics evaluate them. Then 5 speech documents from the ISCI Corpus dataset are selected and transcribed using the Automatic Speech Recognition from Google Cloud Speech API. Now apply extractive summarization algorithms to these 5 text samples to get text summary from the original audio file.

A paper used the Chinese Dataset (SPORTSSUM) **[4]** for generating sports news from live commentary. In it, the

authors have proposed two metrics to evaluate the accuracy of the generated summaries. According to the Authors, the results showed that the proposed model performs well on ROUGE scores and also the two designed scores i.e., name matching score and event matching score. To enhance the quality of generated news, the authors trained the model in a tem2tem way. Another paper presents the use of Heterogeneous Graph Neural Networks **[5]** for summarization. It contains an explanation for the use of semantic nodes of different coherent levels apart from sentences. These nodes act as intermediaries between sentences and thus enrich the cross-sentence relationships. After nodes are created, an extractive text summarization approach is applied to create the summary. There is a paper proposing the Biased TextRank **[6]**. It is graph-based, faster, resource-efficient, language-agnostic, and easy to implement algorithm for extraction of content from the text. The author demonstrated its effectiveness on two tasks: focused summarization and explanation extraction. In 2020, a research paper was published which explained the use of Hierarchical Attentive Heterogeneous Graph Network **[7]**. This paper mainly strives to reduce redundancy created in other approaches like the BERT model. It proposes the use of an extra layer in the neural network called the redundancy layer along with the ALBERT model which is already trained based on similar architecture as that of BERT.

In the paper proposing the Paraphrase Generation **[8]**, according to the authors the Paraphrase generation problem is resolved using conditional generation-based neural networks. The paper proposed a model that performs both tasks by training a single model with the objective of Paraphrase generation. A paper was published in 2021 which focuses mainly on 2 algorithms Textrank and BERT **[9]**. These 2 algorithms are tested using various parameters to get results which one is better than human-generated summaries on news dataset. When results are evaluated using ROUGE scores then it is found that Textrank had a better ROUGE score as compared to BERT. In 2021 a paper explaining the application of the topic modeling approach **[10]** on the WikiHow dataset was published. At the initial step, topics from the input text document are identified using topic modeling techniques like LDA. Then clusters are generated using the topics. Then clusters having salient features are combined to generate a summary. A semantic approach that applies sentence expansion for tuning of conceptual densities **[11]** is also proving effective for automatic text summarization. In it, a framework is proposed which expands each sentence using an innovative approach to reduce ambiguity in the meaning of the sentences and tries to give it meaning which is close to the central topic. Datasets used here are DUC-2002 and DUC-2006. The ROGUE metric is used for output analysis.

6. DETAILED COMPARISON OF RESEARCH PAPERS:

Table-1 Table of Comparison of research papers

Paper-ID	Paper	Dataset	Input Document size	Domain	Approach	Advantage	Disadvantage/ Scope for improvement	ROUGE-1 score	ROUGE-2 score	ROUGE-L score
1	Minakshi Tomer, Manoj Kumar (2021)	DUC-2002, DUC2003, DUC2004	Single	General	Hybrid	1. Higher ROUGE-1, ROUGE-2 scores than the other nature-inspired and swarm-based algorithms	1. Unique fitness functions can be introduced to increase the quality of the summary generated. 2. In the future, it can be used with deep neural based	0.4782	0.2295	0.3362

							models for abstractive text summarization.			
2	Ujjwal Rani and Karambir Bidhan (2021)	Reviews of documents, news articles, legal text.	Single	General	Extractive	1. TextRank is better than TF-IDF and LDA.	1. Other approaches for text summarization are available which give best results.	Review dataset - f-measure 0.2330 News dataset - f-measure 0.6367 Legal dataset - f-measure 0.26	Review dataset - f-measure 0.0550 News dataset - f-measure 0.6139 Legal dataset - f-measure 0.0960.	Review dataset - f-measure 0.1798 News dataset - f-measure 0.6520 Legal dataset - f-measure 0.2346.
3	Begum Mutlu, Ebru A. Sezer, M. Ali Akcayol (2019)	DUC 2002 dataset	Multi-document	General	Extractive	1. It shrinks the main data size so that the summary achieved as a result replaces the initial document.	1. The Computation of power for NLP is utilized on a large scale by the abstraction, then parsed and generated after the grammar are included along with lexicons.	0.398	0.085	Unknown

4	Kuan-Hao Huang, Chen Li, Kai-Wei Chang (2020)	SPORTSSUM	Multi document	Domain specific-News domain	Hybrid	1. In this paper, two metrics were designed by the authors to assess the correctness of generated summaries of live sports commentary. 2. And, according to the authors, results show that the proposed model performs well on ROUGE scores and the two designed scores or metrics.	1. More research on the tem2tem approach is required for improving the correctness of summaries.	0.244	0.063	0.231
5	Danqing Wang, Pengfei Liu, Yining Zheg, Xipeng Qiu, Xuanjing Huangn (2020)	CNN/Dailymail, NYT50 and Multi News	Single-document and multi-document	Domain Specific-News domain	Hybrid	1. This method proves very much effective for multi-document summarization which needs to maintain the inter-document relation to produce accurate and effective text summarization. 2. It also provides best results as compared to other non-BERT based models.	1. In future, pre-trained language models can also be considered for increasing the nodes' encoding representation.	CNN/DailyMail: 0.4295, NYT50: 0.4389, Multi News: 0.4605	CNN/Daily Mail: 0.1976, NYT50: 0.2626, Multi News: 0.1625	CNN/Daily Mail: 0.3923, NYT50: 0.4258, Multi News: 0.4208
6	Ashkan Kazemi, Veronica Perez-Rosas, Rada Mihalcea (2020)	novel dataset	Single	General	Extractive	1. Biased TextRank is easy to implement, it is faster, lighter than current state-of-the-art Natural Language Processing methods for similar tasks.	1. In future there's a scope to explore the applications of Biased TextRank beyond sentence extraction	Democrat - 0.3009, Republican - 0.3366	Democrat - 0.0584, Republican - 0.0585	Democrat - 0.2135, Republican - 0.2211

7	Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang1, Cong Cao and Shi Wang (2020)	CNN/DailyMail, NYT, Newsroom (Ext)	Single	Domain specific-News domain	Hybrid	1. The extra neural network layer of redundancy added in the ALBERT proves very useful in reducing the redundancy which unnecessarily increases the length of the text summary.	1. This model is relatively harder to understand and implement	CNN/DailyMail: 0.4468, NYT: 0.4936, Newsroom: 0.7131	CNN/Daily Mail: 0.2130, NYT50: 0.3141, Multi News: 0.6875	CNN/Daily Mail: 0.4075, NYT50: 0.4497, Multi News: 0.7083
8	Hemant Palivelaa (2021)	ParaNMT	Single	General	Hybrid	1. In this paper, the authors proposed a model that can perform both tasks like training a single model with the objective of paraphrase generation.	1. In this paper, the T5-base model is calibrated perfectly but it can be expanded to T5-large and other variants too.	0.52	0.35	0.5
9	Sreeya Reddy Kotrakona Harinatha, Beauty Tatenda Tasara , Nunung Nurul Qomariyah (2021)	News dataset	Single	Domain specific-News article	Extractive	1. TextRank has a better ROUGE score as compared to BERT. TextRank showed higher F-measure and recall.	1. BERT has higher precision than textrank.	0.6004	0.5892	0.5668
10	Kalliath Abdul Rasheed Issam, Shivam Patel, Subalalitha C. N. (2021)	WikiHow articles	Single	Domain specific-WikiHow articles	Extractive	1. Since the dataset used are very short abstract text summaries, the model performed really well and thus ensures that it can perform brilliantly when provided with an appropriate dataset.	1. The paper doesn't mention anything related to multiple document text summarization	0.2708	0.0689	0.254

11	Mohammad Bidoki, Mohammad R. Moosavi, Mostafa Fakhrahmad(2020)	DUC-2002 and DUC-2006	Multi-document	General	Hybrid	1. The model is language independent. 2. It dynamically recognizes the clusters and extracts them to create the summary.	1. As part of future development, there can be improvement in the redundancy reduction, readability, etc.	DUC-2002: 0.5137, DUC-2006: 0.4053	DUC-2002: 0.2605, DUC-2006: 0.1126	-

7. CONCLUSION

In this paper, a detailed comparative study of available automatic text summarization techniques has been presented. It is observed that the ATS techniques can be evaluated fairly based on their ROUGE scores (more specifically their ROUGE-1, ROUGE-2, and ROUGE-L scores). The most ideal datasets considered for testing and training of the text summarization models include the Document understanding Conference (DUC) datasets, CNN/DailyMail, NYT50, etc. The ROUGE-1 score above 0.4, ROUGE-2 score above 0.15, and ROUGE-L score above 0.33 are very promising to give a perfect summary of the text document. However, there is a dearth of standard datasets for generating summaries from sports commentary. Generating a benchmark dataset for evaluating summaries generated from sports commentary can be useful work for future researchers.

The major problem in text summarization is that of the redundancy which leads to unnecessary lengthening of the summary. Most of the papers proposed models which worked on single document text summarization but not on multi-document text summarization. Very few papers are identified which can handle large size text documents and still produce remarkable results. Some papers proposed a solution for specific genres like business, sports, news, etc. but can be used in other genres with less promising results. Making these solutions generic and improving the results thereafter can be another challenging task.

Some research papers proposed models which were language specific. Improvement can be done to make the model diverse in terms of the language input. As most of the models are in the primitive stage, progress can be made to increase the reliability and readability of the summary generated from the models.

Different research and experiments have been carried out for overcoming the problems faced in text summarization and exploration is done to find new solutions for long. Due to the ever-increasing abundance of textual data, it is a need to create summaries and focus on the required data. In earlier stages, the focus was more on extractive summarization. Because of the introduction of abstractive summarization, there was a rise in various solutions of a given problem with more efficient summary generation.

Still, many challenges need to be addressed in the field of automatic text summarization. A major challenge in text summarization is the redundancy in the document. Though some solutions are proposed to reduce redundancy those are not very impressive, and redundancy remains a hurdle in generating precise summaries. More work is needed in maintaining the continuity of the generated summary for a single document as well as multi-document text summarization. The challenge of continuity lies more in multi-document text summary as compared to single document text summary. For future work, the models can be improved for the diversity of input documents while maintaining the precision of generating a summary.

From the discussion above, it is understood that there lies lots of research gaps in the domain of automatic text summarization and the existing approaches need improvement as well. It is hoped that this study will help researchers diminishing those identified research gaps in various areas of automatic text summarization.

REFERENCES

[1] Tomer, Minakshi, and Manoj Kumar. "Multi-document extractive text summarization based on firefly algorithm." *Journal of King Saud University-Computer and Information Sciences* (2021).

[2] Rani, Ujjwal, and Karambir Bidhan. "Comparative assessment of extractive summarization: textrank tf-idf and lda." *Journal of Scientific Research* 65, no. 1 (2021): 304-311.

[3] Mutlu, Begum, Ebru A. Sezer, and M. Ali Akcayol. "Multi-document extractive text summarization: A comparative assessment on features." *Knowledge-Based Systems* 183 (2019): 104848.

[4] Huang, Kuan-Hao, Chen Li, and Kai-Wei Chang. "Generating Sports News from Live Commentary: A Chinese Dataset for Sports Game Summarization." In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 609-615. 2020.

[5] Wang, Danqing, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. "Heterogeneous graph neural networks for extractive document summarization." *arXiv preprint arXiv:2004.12393* (2020).

[6] Kazemi, Ashkan, Verónica Pérez-Rosas, and Rada Mihalcea. "Biased TextRank: Unsupervised graph-based content extraction." *arXiv preprint arXiv:2011.01026* (2020).

[7] Jia, Ruipeng, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. "Neural extractive summarization with hierarchical attentive heterogeneous graph network." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3622-3631. 2020.

[8] Palivela, Hemant. "Optimization of paraphrase generation and identification using language models in natural language processing." *International Journal of Information Management Data Insights* 1, no. 2 (2021): 100025.

[9] Harinatha, Sreeya Reddy Kotrakona, Beauty Tatenda Tasara, and Nunung Nurul Qomariyah. "Evaluating Extractive Summarization Techniques on News Articles." In *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pp. 88-94. IEEE, 2021.

[10] Issam, Kalliath Abdul Rasheed, and Shivam Patel. "Topic Modeling Based Extractive Text Summarization." *arXiv preprint arXiv:2106.15313* (2021).

[11] Bidoki, Mohammad, Mohammad R. Moosavi, and Mostafa Fakhrahmad. "A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities." *Information Processing & Management* 57, no. 6 (2020): 102341.