Briefing of Textual Information Using TextRank

Dr.G.N Beena Bethel
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
beenabethel@gmail.com

Baki Divyasree
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
dsbaki05@gmail.com

Vangapandu Sandhya
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
vangapandusandhya@gmail.com

K A Nimisha
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
kotagirinimisha@gmail.com

Anjani Sreemanth Bodduluri
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
Sreemanth.bodduluri@gmail.com

Abstract -- The world is very much advanced with the abundant growth of technology and the way we communicate is rapidly changing with it. Lot of information is being generated day by day. It is important to extract the useful information from it. Text Summarization is one of the solutions to it. Different statistical methods like TFIDF, TF are used for extracting the summary. These statistical methods focus mainly on most frequently occurred words and less on importance of sentences. There are many limitations of TFIDF method like it is based on bag of words model, so it does not capture position in text, semantics, co-occurrences in different documents. The other limitation is it assigns low values to words that are relatively important. So, we proposed a system that uses a graph-based approach which extracts the significant sentences from the given text. The proposed system provides Extractive Summary, Abstractive Summary and Keywords for single document. The keywords extracted from the summary helps in understanding the main idea of the document. The project that we proposed helps users in providing summary with important points and saves user time.

Keywords -- Text Summarization, Single Document Summarization, Multi Document Summarization, Keywords Extraction, TFIDF, Text Rank.

I. INTRODUCTION

In the present generation a lot of data is generating through the internet. It is important to provide the best method for generating significant information fast and effectively. One of the best methods is Text Summarization. It helps us in reducing the time required for reading whole document and, we can reduce the space required to store large amount of data. Text Summarization can be classified into different types [1] as shown in "Fig.1". Based on input document it can be divided into two types: single document summarization [5] and multi document summarization [13]. Based on language it is classified into Monolingual, Multilingual [14] and Cross language. Based on types of approaches it can be classified into Extractive Summarization [15] and Abstractive Summarization [16]. An Extractive text summarization is generated by selecting the important lines from the given text. Extractive Summarization can be done using two phases: pre-processing and post-processing.

An Abstractive Text Summarization will try to understand the input file and regenerate the summary by identifying key points in the text. Keywords Extraction [9] from the summary helps in understanding the main idea of document.

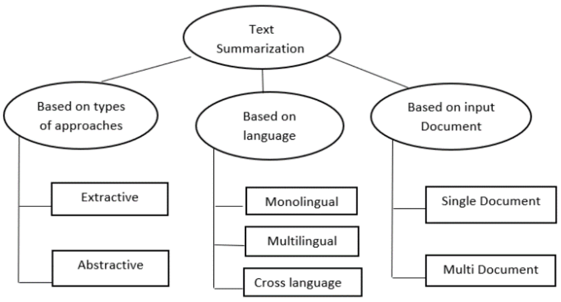


Fig 1: Text Summarization Techniques

This proposed paper explains Extractive summarization, Abstractive summarization, Keywords Extraction on single document. We are using graph-based method for extracting summary. We are also comparing the performances of TFIDF method and graph-based method [17]. Term Frequency Inverse Document Frequency [5] works by considering the frequency of each term in a sentence. Our system used graph-based method which represents the connections between the objects. In text documents we can refer vertex to sentences and edges to weights between two sentences. Our system is proved in providing accurate and efficient summary.

II. LITERATURE SURVEY

The authors of paper [1] has explained different types of text summarization techniques and different methods of extractive summarization and abstractive summarization.

The authors of paper [11] used sentence ranking method for extractive summarization. The weights are assigned to each sentence and ranked based on their weights.

The authors of paper [15] used the k-means clustering and TFIDF method for extracting summary and compared both the methods. For calculating K value in K-means clustering the Elbow method and Silhouette method can be used.

The authors of paper [5] has proposed TFIDF for Single Document Summarization. He compared his proposed system with three other different online summarizers. Here the document has taken and manually the summary is generated by a group of people. The comparison has taken based on-line numbers of sentences that are generated by human and proposed system. The evaluation is based on three parameters: Precision, Recall, and F-measure. The accuracy obtained is 67%.

The authors of paper [2] used Natural Language Tool Kit (NLTK) and Spacy module for generating summary. A webpage is developed so that user can provide the input text and the summary is provided as output.

III. METHODOLOGY

In our paper we designed a user-friendly Graphical User Interface (GUI) which allows the user to access the following modules.

A. Extractive Summarization

In this, the summary is generated by including the most important parts of the sentences in the given user input text [15]. The proposed system uses sentence vectorization, graph-based method called Text Rank algorithm, and cosine similarity to check the document similarity.

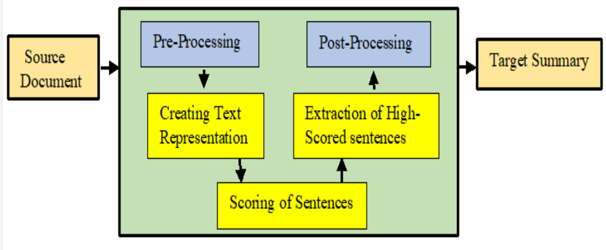


Fig 2: Flow-Chart of Extractive based Summarization.

B. Keywords Extraction

On selecting this tab, the keywords are displayed to user by going through the whole set of data and obtain the words and phrases that best describe each context.

C. Abstractive Summarization

It generates summary that involves the words which are generally not present in the actual text [16]. It is the technique of generating a summary of a given text from its hidden meaning, but not by framing a new sentence by extracting the main keywords from the text.

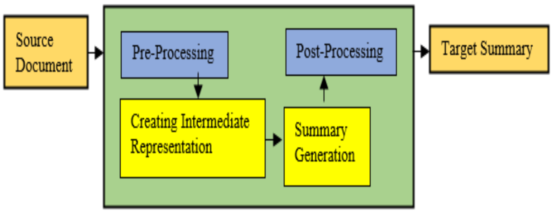


Fig 3: Flow-Chart of Abstractive based Summarization.

IV. IMPLEMENTATION

A. Extractive Summarization

The steps in this summarization are:

1. Pre-processing:

- Tokenization: It is the process of breaking down the sentences into words called units to maintain the similarity.
- Lower Case Conversion: The process of changing the case of all tokens to avoid redundancy and ambiguity.
- Stop words removal: The words which are responsible for noise are removed in this step.
- Lemmatization: The base form of the word is extracted here are called as lemma.
- Parse tree generation: For extracting all parts of speech from the sentences we can use NLTK RegexpParser.
- POS Tagging: Parts of Speech tagging should be done in order to avoid the confusion between two same words that have different meanings.

2. Text Rank algorithm:

It is one of the graph-based approach that is used to find most important sentences and keywords from the given text [19]. Firstly, it concatenates all the text that is present in the articles. The text is then split to get individual sentences. Every sentence is represented by a vector. Then the cosine similarity is used to check the similarities between sentence vectors and the values are stored in a matrix.

$$\boxed{\text{Similarity}(C,D)=\|C\|\,\|D\|\cos(\text{theta})=C.D}$$

The similarity matrix is then converted into a graph. Here the vertices represent sentences and edges represents similarity scores.

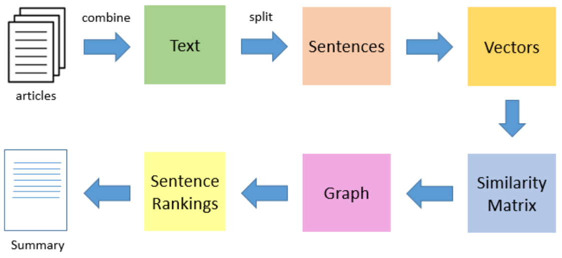


Fig 4: Flow-Chart of Text Rank Algorithm.

3. Post-Processing:

In post processing step the summary is produced by selecting k highly ranked sentences.

B. Keywords Extraction

In this, our system displays the keywords by considering the words that are mostly used in framing a sentence which are to be included in the generating summary.

C. Abstractive Summarization

The abstractive based summary is obtained by undergoing the following steps.

Firstly, we perform text pre-processing and tokenization. Next the encoder encodes the given input text and passes it to the decoder which produces the summary of text by selecting the representation from the final states.

V. RESULTS

We evaluated our proposed system using the Kaggle dataset consisting of 2,225 BBC articles which belong to entertainment, business, politics, sports, and technology. From the dataset we took 10 news articles of entertainment as input. For every input document our proposed system generated summary which was about 50% of the original document. We compared our proposed system generated summaries with the MS-Word generated summaries and with TF-IDF method generated summaries. According to [18], the evaluation process of text summarization is performed by using three parameters which are precision, recall, and f-measure.

The evaluation metrics for all the three systems were calculated with respect to the human self-obtained summaries. We have asked few people to generate their respective summaries for the given input articles that were used for comparison of all the three systems.

Table 1: Sentences included in the summary generated by MS-Word Summarizer, Proposed System Summarizer, Human from its corresponding input document.

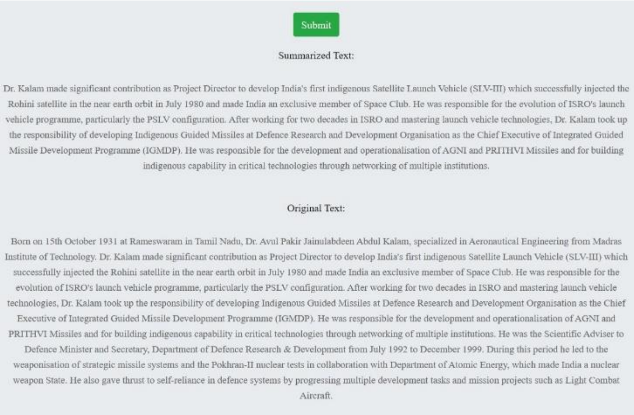
Document	Total	MS-Word	TF-IDF	Proposed system	Human
1	13	1,2,3,4,5,6,7	1,2,3,5,6,7,8	1,2,3,4,7,8	1,2,3,4,9,11
2	12	1,2,3,4,6,8,12	1,2,6,9,10	1,2,3,4,8	1,2,3,4,8,11
3	12	1,2,4,5,6,7,11	1,2,3,4,7,11,12	1,2,3,4,7,11,12	1,2,5,8,11,12
4	14	1,2,3,4,5,6,8,14	1,4,5,6,11,13	1,4,5,6,11,13	1,4,5,6,7,8,10,11,12,14
5	12	1,5,6,8,9,11,12	1,4,5,6,7,11	1,2,6,7,9,11,12	1,2,6,7,8,9,10
6	10	1,2,5,6,7,9	1,4,5,8,9,10	1,2,6,9,10	1,2,6,9
7	13	1,2,5,8,10,12	1,6,9,11,12,13	1,2,5,7,11,12,10	1,2,5,7,8,11,13
8	11	1,2,3,4,5,6	1,5,6,8,9,10,11	1,2,4,6,9,10,11	1,2,4,6,7,9
9	11	1,2,3,4,6,11	4,5,6,9	4,5,6,9	2,3,4,5,6,7,10
10	11	1,2,3,5,7,8	1,4,5,6,9,10	2,3,5,6,7	2,3,5,6,7,9

Table 2: MS-Word, TF-IDF and proposed system’s summarizer evaluation measures.

Document	MS-Word Evaluation			TF-IDF Evaluation			Proposed Evaluation		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
1	0.571	0.667	0.615	0.8	0.667	0.727	0.667	0.667	0.667
2	0.714	0.833	0.769	0.571	0.667	0.615	1.000	0.833	0.908
3	0.571	0.667	0.615	0.571	0.667	0.615	0.714	0.833	0.769
4	0.750	0.600	0.833	0.833	0.500	0.624	1.000	0.714	0.823
5	0.571	0.571	0.714	0.714	0.714	0.714	1.000	0.857	0.923
6	0.667	1.000	0.333	0.333	0.500	0.4	0.250	0.250	0.250
7	0.667	0.571	0.714	0.714	0.714	0.714	0.857	0.857	0.857
8	0.667	0.667	0.667	0.714	0.833	0.768	0.800	0.667	0.727
9	0.667	0.571	0.615	0.750	0.42	0.538	1.000	0.857	0.922
10	0.667	0.667	0.667	0.500	0.500	0.500	0.833	0.833	0.833
Average(%)	65	68	66	67	61	62	81	73	77

Table [1] displays the number value of sentence which is included in the summary. Table [2] displays the precision, recall and f-measure of our proposed system along with MS-Word and TF-IDF. The objective of our proposed system is to provide the complete usability of our graphical user interface to the user. From the graphical user interface, the user is free to choose the type of summarizer to obtain the summary of given input text. The user can also extract the important keywords of the given text.

The below figures are the screen shots of our system generated summarized text and key words extracted for the given original text.



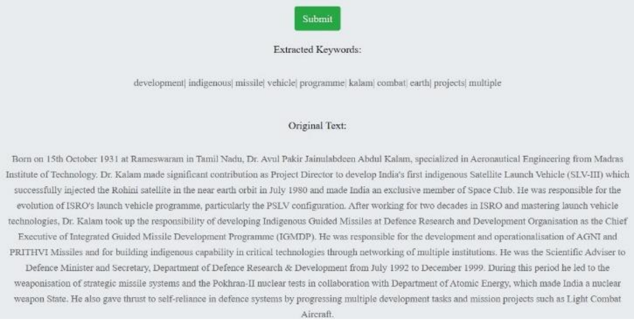
Summarized Text:

Dr. Kalam made significant contribution as Project Director to develop India's first indigenous Satellite Launch Vehicle (SLV-III) which successfully injected the Rohini satellite in the near earth orbit in July 1980 and made India an exclusive member of Space Club. He was responsible for the evolution of ISRO's launch vehicle programme, particularly the PSLV configuration. After working for two decades in ISRO and mastering launch vehicle technologies, Dr. Kalam took up the responsibility of developing Indigenous Guided Missiles at Defence Research and Development Organisation as the Chief Executive of Integrated Guided Missile Development Programme (IGMDP). He was responsible for the development and operationalisation of AGNI and PRITHVI Missiles and for building indigenous capability in critical technologies through networking of multiple institutions.

Original Text:

Born on 15th October 1931 at Rameswaram in Tamil Nadu, Dr. Avul Pakir Jainulabdeen Abdul Kalam, specialized in Aeronautical Engineering from Madras Institute of Technology. Dr. Kalam made significant contribution as Project Director to develop India's first indigenous Satellite Launch Vehicle (SLV-III) which successfully injected the Rohini satellite in the near earth orbit in July 1980 and made India an exclusive member of Space Club. He was responsible for the evolution of ISRO's launch vehicle programme, particularly the PSLV configuration. After working for two decades in ISRO and mastering launch vehicle technologies, Dr. Kalam took up the responsibility of developing Indigenous Guided Missiles at Defence Research and Development Organisation as the Chief Executive of Integrated Guided Missile Development Programme (IGMDP). He was responsible for the development and operationalisation of AGNI and PRITHVI Missiles and for building indigenous capability in critical technologies through networking of multiple institutions. He was the Scientific Adviser to Defence Minister and Secretary, Department of Defence Research & Development from July 1992 to December 1999. During this period he led to the weaponisation of strategic missile systems and the Pokhran-II nuclear tests in collaboration with Department of Atomic Energy, which made India a nuclear weapon State. He also gave thrust to self-reliance in defence systems by progressing multiple development tasks and mission projects such as Light Combat Aircraft.

Fig 5: Extractive summary generated for the original text.



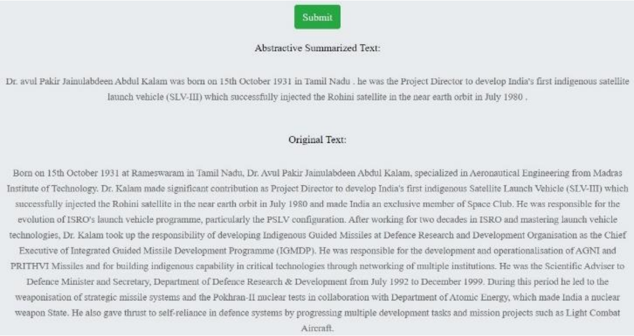
Extracted Keywords:

development| indigenous| missile| vehicle| programme| kalam| combat| earth| projects| multiple

Original Text:

Born on 15th October 1931 at Rameswaram in Tamil Nadu, Dr. Avul Pakir Jainulabdeen Abdul Kalam, specialized in Aeronautical Engineering from Madras Institute of Technology. Dr. Kalam made significant contribution as Project Director to develop India's first indigenous Satellite Launch Vehicle (SLV-III) which successfully injected the Rohini satellite in the near earth orbit in July 1980 and made India an exclusive member of Space Club. He was responsible for the evolution of ISRO's launch vehicle programme, particularly the PSLV configuration. After working for two decades in ISRO and mastering launch vehicle technologies, Dr. Kalam took up the responsibility of developing Indigenous Guided Missiles at Defence Research and Development Organisation as the Chief Executive of Integrated Guided Missile Development Programme (IGMDP). He was responsible for the development and operationalisation of AGNI and PRITHVI Missiles and for building indigenous capability in critical technologies through networking of multiple institutions. He was the Scientific Adviser to Defence Minister and Secretary, Department of Defence Research & Development from July 1992 to December 1999. During this period he led to the weaponisation of strategic missile systems and the Pokhran-II nuclear tests in collaboration with Department of Atomic Energy, which made India a nuclear weapon State. He also gave thrust to self-reliance in defence systems by progressing multiple development tasks and mission projects such as Light Combat Aircraft.

Fig 6: Keywords generated for original text.



Abstractive Summarized Text:

Dr. avul Pakir Jainulabdeen Abdul Kalam was born on 15th October 1931 in Tamil Nadu . he was the Project Director to develop India's first indigenous satellite launch vehicle (SLV-III) which successfully injected the Rohini satellite in the near earth orbit in July 1980 .

Original Text:

Born on 15th October 1931 at Rameswaram in Tamil Nadu, Dr. Avul Pakir Jainulabdeen Abdul Kalam, specialized in Aeronautical Engineering from Madras Institute of Technology. Dr. Kalam made significant contribution as Project Director to develop India's first indigenous Satellite Launch Vehicle (SLV-III) which successfully injected the Rohini satellite in the near earth orbit in July 1980 and made India an exclusive member of Space Club. He was responsible for the evolution of ISRO's launch vehicle programme, particularly the PSLV configuration. After working for two decades in ISRO and mastering launch vehicle technologies, Dr. Kalam took up the responsibility of developing Indigenous Guided Missiles at Defence Research and Development Organisation as the Chief Executive of Integrated Guided Missile Development Programme (IGMDP). He was responsible for the development and operationalisation of AGNI and PRITHVI Missiles and for building indigenous capability in critical technologies through networking of multiple institutions. He was the Scientific Adviser to Defence Minister and Secretary, Department of Defence Research & Development from July 1992 to December 1999. During this period he led to the weaponisation of strategic missile systems and the Pokhran-II nuclear tests in collaboration with Department of Atomic Energy, which made India a nuclear weapon State. He also gave thrust to self-reliance in defence systems by progressing multiple development tasks and mission projects such as Light Combat Aircraft.

Fig 7: Abstractive summary generated for original text.

VI. CONCLUSION AND FUTURE DEVELOPMENT

Hence from the above results it is observed that the precision, recall and f-measure values of the proposed system generated summaries are higher than the MS-Word generated summaries and TFIDF. The average F-Measure of our proposed system is 77% which is greater than 66% of MS-Word and 62% of TF-IDF system. Our system works well as it does not depend on single vertex but works recursively on obtaining the information from entire graph. Our system is proved in providing the accurate summary which is efficient and time saving. In the future we will expand our system to generate summaries for multiple documents and also to improve the performance of abstractive based summarizer along with using different evaluation measures for its comparison with other systems.

VII. REFERENCES

[1] Vinit Aghama, Dr.V.K.Shandilyab. "A Survey Paper on Extractive and Abstractive Techniques in Automatic Text Summarization", International Journal of Research Publication and Reviews Vol (2) Issue (4) (2021) Page 619-625.

[2] B.Hemanth Kumar, L. Ramaparvathy. "GUI Based Text Summarizing of Social Response", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020.

[3]Yogan Jaya Kumar, Ong Sing Goh, Halizah Basiron, Ngo Hea Choon and Puspalata C Suppiah. "A Review on Automatic Text Summarization Approaches", Journal of Computer Science, 2016.

[4]Yogesh Kumar Meena, Dinesh Gopalani. "Evolutionary Algorithms for Extractive Automatic Text Summarization", Procedia Computer Science , Volume 48, 2015, pages 244-249.

[5]Hans Christian, Mikhael Pramodana Agus, Derwin Suhartono. "SINGLE DOCUMENT AUTOMATIC TEXT SUMMARIZATION USING TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)", ComTech Vol. 7 No. 4 December 2016: 285-294.

[6]Amey Takur, Mega Satish. "Text Summarizer Using Julia", International Journal for Research in Applied Science and Engineering TechnologyISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue I Jan 2022.

[7]Dr. Geetha C Megharaj , Ms. Varsha Jituri. "TFIDF Model based Text Summarization", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Volume 10, 2022.

[8]Saurabh Varade , Ejaaz Sayyed , Vaibhavi Nagtode1, and Dr. Shilpa Shinde1. "Text Summarization using Extractive and Abstractive Methods", EDP Sciences, 2021.

[9]Santosh Kumar Bharti, Korra Sathya Babu, and Anima Pradhan. "Automatic Keyword Extraction for Text Summarization inMulti-document e-Newspapers Articles", European Journal of Advances in Engineering and Technology, 2017.

[10]Jay Sharma, Harsh Hardel, Chirag Sahuji, Rajesh Prasad. "Automatic Text Summarization", International Research Journal of Engineering and Technology, Volume: 09 Issue: 06, June 2022.

[11]J.N Madhuri, R. Ganesh Kumar. “Extractive Text Summarization using Sentence Ranking”. IEEE,2019.

[12] Deepali K. Gaikwad, C. Namrata Mahender. “A review paper on text summarization”. International Journal of Advanced Research in Computer and Communication Engineering, March 2016.

[13]Yong Zhang, Meng Joo Er, Rui Zhao, and Mahardhika Pratama. “Multiview Convolutional Neural Networks for Multidocument Extractive Summarization”, IEEE Transactions on Cybernetics 47, 10, 3230–3242.

[14] Negar Foroutan , Angelika Romanou , Stephane Massonnet, Remi Lebret, Karl Aberer. “Multilingual Text Summarization on Financial Documents”, European Language Resource Association (ELRA), 2022.

[15]Rahim Khan, Yurong Qian, Sajid Naeem. “Extractive based Text Summarization Using K-Means and TF-IDF”, International Journal of Information Engineering and Electronic Business, May 2019.

[16]N. Moratanch, Dr.S.Chitrakala. “A Survey on Abstractive Text Summarization”, International Conference on Circuit, Power and Computing Technologies (ICCPCT), 2016.

[17] Asahi Ushio, Federico Liberatore, Jose Camacho-Collados. “A Quantitative Analysis of Statistical and Graph-Based Term Weighting Schemes for Keyword Extraction”, Online Research @ Cardiff, 2021.

[18]Nedunchelian, R., Muthucumarasamy, R., & Saranathan, E. (2011). Comparison of multi document summarization techniques. International Journal of Computer Applications. 11(3), 155-160.

[19]Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the 42nd Annual Meeting of the Association for Computational Lingusitics (ACL 2004) (companion volume), Barcelona, Spain.