Automated Script Outline Generator using TextRank Algorithm

Urvi Zope [1], *Gayatri Thakur* [2], *Siddhi Karambelkar*[3]*and Vanita Mane*[4]

Department of Computer Engineering, Ramrao Adik Institute of Technology, Dr. D. Y. Patil Deemed to be University, Nerul, Navi Mumbai, Maharashtra, India[1,2,3,4]

Abstract

The process of creating a condensed type of document that retains significant information and thus the general meaning of the source text is known as text summarization. Automatic summarization is a popular method for quickly locating significant information in vast texts with minimal effort. In today's era information is the integral part of knowledge and it is too noisy and not in the precise manner for that both abstractive and extractive summarization techniques are used. The proposed model combines the summarization process for both text and image input type, which makes it easier for user to generate the precise summary regardless of the input type. The proposed model uses Textrank algorithm for the summarization process. Recall oriented understudy of gisting evaluation (ROUGH) score is used by the model to evaluate the accuracy of generated output. The model gives high ROUGH score which assures the accuracy of the generated summary.

1. Introduction

Automatic summarization technology is critical for information retrieval and text classification, and it should give a solution to the problem of knowledge overload [1]. A procedure of lowering the size of a text while keeping its information content intact is known as text summarization.[2] The proposed model uses a sentence clustering-based summarization approach. The model uses textrank algorithm for summarization which works on sentence ranking method. The model compares its results with researches in the field to provide more efficient summary of the data. The proposed model combines the text and image summarization model in a single system which differentiate it from the previous generated models.

1.1 Objectives

This model consists various input type summarization in an integrated form. This model aims at accurate summarization of provided input data. The generated output will be easily understandable and readable. Proposed system will analyze the given input data and remove the irrelevant information and will integrate the central ideas in meaningful way thus the model will focus on the main corpus of information rather than the information which leads to information overload. It will generate a compact yet precise output. Effective text summarization: The model aims to provide the best solution for summarization of the given input data regardless of its type. The proposed system aims to summarize the data in an accurate way regardless of the size of input data.

1.2 Motivation

In this era of information technology information plays a vital role in everyday life of all the individuals.[3] The summarization makes the process of information gathering easy for the users. The proposed model compared its output with the output of previous systems which helped to increase the usability and provide better result to end users of the system.

2. Literature Survey

2.1 Existing Systems

The literature survey for the summarization is briefly explained in the following segment. The main discussion is associated with various algorithms used for summarization and the output generated by all the algorithms respectively.

A. Term Frequency (TF) - Inverse Document Frequency (IDF): TFIDF is the algorithm used in extractive summarization. The terminology Term Frequency (TF) is used to count the occurrence of the words.
The frequency obtained is used to determine the word's importance. The greater the occurrence of a word, the greater its significance in the paper [4]. The simplest way to explain TF is that it counts the number of times a word appears in a document [5]. IDF stands for Inverse Document Frequency, which gives unusual words a greater value and recurrent terms a smaller one. TF sometimes miscalculates the importance of stop words based on their frequency of occurrence. IDF identifies the infrequent occurrence of words in the document to solve TF's problem.

B. Sequence to Sequence Model: The suggested TFRSP approach uses an abstractive summarization algorithm known as a sequence-to-sequence model to create new phrases while keeping the sense of the source content. Google was the first to introduce the Sequence-to-Sequence methodology, which now drives applications such as Google Translate, image captioning, text summarization, online chat bots, and more. It is an encoder-decoder paradigm that maps input and output sequences of various lengths to each other [6]. Long Short-Term Memory (LSTM) is a subcomponent of the encoder-decoder component that is effective for capturing long-term dependencies. There are two steps to the encoder-decoder model: training and inference [7].

C. Text Rank Algorithm: For summarization, the Text Rank algorithm is an extractive and unsupervised learning system. This algorithm is based on the methodology of sentence ranking. As a result, the data in a given input file is first concatenated together in the Text Rank method. After that, stemming is used to separate it into independent sentences [8]. Then, for each sentence, word embeddings are calculated. After that, the algorithm determines sentence-to-sentence similarity and saves the results in a matrix. The similarity matrix is then transformed into a graph, with the sentences serving as vertices and the similarity matrix serving as edges [6]. This graph is then used to rank the sentences, which aids in the creation of the summary. The sentences with the highest scores are given out first.

D. ROUGE score: The ROUGE metric (Recall Oriented Understudy of Gisting Evaluation) is a text summarization metric. It's used to compare the n-gram matches between the system-summary acquired from text summarization and the human-generated reference summary [9]. The ROUGE score is made up of precision, recall, and f-measure values that are added together. The ROUGE-
1 score assesses the overlap of unigrams from the system-generated summary and the reference summary created by humans [10].

2.2 Limitations of Existing System

The existing systems have given good results. But they do not have all the required features. Basically, they are either text input dependent or image input dependent. Following research gaps are identified in the existing techniques:

- Existing model provides a correct summary of only short text. When it came to longer text the summarization performed by these systems was inaccurate and irrelevant.
- The existing systems developed for the summarization were unable to understand the meaning of many vocabulary words. Some of the systems are tested only for specific language datasets. So, the accuracy of generated abstract is not that good when it came to the vocabulary words.
- Many difficulties were faced by the systems while summarizing the long texts.
- As discussed before previously designed systems were unable to generate accurate abstract for the loner texts. It is also observed that when the number of selected sentences exceeds three, the average ROUGE score drops.
- There is no existing model which combined both text and image summarization model into single system.

3. Proposed System

3.1 Problem Statement

"***Automated Script Outline Generator using textrank Algorithm***" is a model that aims to build a system which will be able to perform the summarization of both text and image. Based on the given input the proposed system will extract the important content from that.

3.2 Proposed Methodology

The proposed model uses textrank algorithm for the summarization of given input data. The following diagram shows the detailed working of the proposed system. As shown in the fig. 1 The model first takes the input from the user in either text or image format. If the given input is in image format, then it is first converted into text using OCR. Then Document preprocessing is performed in which stemming and lemmatization is performed on the input data. Then all the data in collected together then textrank algorithm is applied onto that data. The sentences are ranked in a tree manner. User defines the amount of summary required by them. Accordingly, system generates summary with highest ranked sentences. In this way the proposed model works. The ranking algorithms thus provides correctness to the overall output as only the highest ranked sentences are given out as output summary.

The practice of distilling the most significant information from a source into an abridged version for a specific user and task is known as summarizing. A sentence clustering-based summarization strategy is proposed in this model. A system pipeline that utilizes both error detection and correction models.

The summarization process will be performed using textrank algorithm where first the system will take the input data, perform pre-processing on that data and then then model building will be done after which the summarized output data will be generated.

3.2.1 Input:

Input consists of the input data that user is going to provide to the system. For this proposed system BBC news dataset from Kaggle is used. This dataset is made by collecting various news articles gathered from all over the web. Also for the image input the system is provided with an image containing information about android. Thus the system is using the multiple sentence dataset for the evaluation purpose.

3.2.2 Data Pre-processing:

In the preprocessing part of implementation, the system performs stemming and lemmatization on the given input data. In this module first the system will Read the file**.** After that it will Split the text in the articles into sentences. Then the system will remove punctuations, numbers and special characters. The system will also remove stop words from the sentences. After all this procedure the system will extract word vectors.

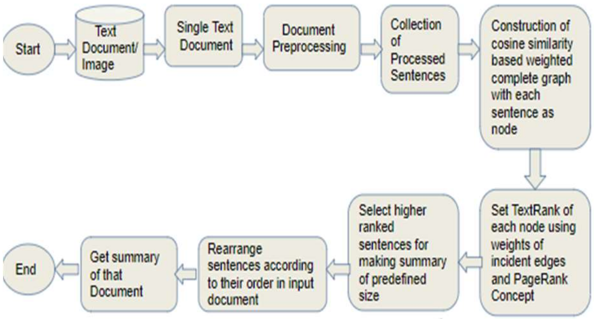

Start
Text Document/ Image
Single Text Document
Document Preprocessing
Collection of Processed Sentences
Construction of cosine similarity based weighted complete graph with each sentence as node
Set TextRank of each node using weights of incident edges and PageRank Concept
Select higher ranked sentences for making summary of predefined size
Rearrange sentences according to their order in input document
Get summary of that Document
End

Fig. 1 Block diagram of automated script outline generator

3.3.3. Model Building using TextRank algorithm:

After the preprocessing is performed on the data the next step in the model is model building. In this step basically each sentence is then tokenized into a group of words in the next phase. Then, in order to define sentence similarity, we represent each sentence in the given text document as a word vector. Similarities between sentence vectors are calculated and stored in a matrix using cosine similarity. The similarity matrix is then turned into a graph for sentence rank calculation, with sentences as vertices and similarity scores as edges. Finally, the final summary is made up of a set of top- ranked sentences.

Text Rank algorithm is extractive and unsupervised learning algorithm used for the summarization. This algorithm works on sentence ranking methodology. So, in Text Rank algorithm first of all the data in given input file is concatenated together. Then it is splitted into individual sentences by performing stemming on it. After that word embeddings for each sentence are calculated. Then the algorithm calculates the sentence-to-sentence similarity and stores the calculated similarities in matrix. The similarity matrix is then converted to form a graph where the

https://doi.org/10.1051/itmconf/20224403035

sentences act as vertices of graph and the similarity matrix acts as edges of graph. This graph is then used to rank the sentences this ranked sentences helps to generate the summary. The highly ranked sentences are given out as the generated output summary.

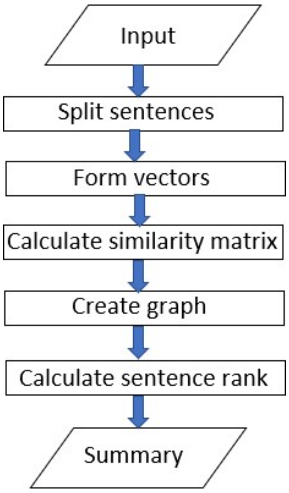


Fig. 2 TextRank algorithm

4. Result and Analysis

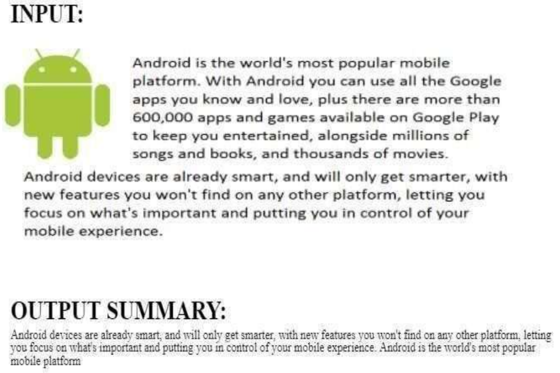
4.1 Evaluation parameters

The generated output by the proposed model is compared with other existing summarization models using ROUGH score. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is an acronym for ROUGE. It consists mostly of a set of measures for assessing automatic text summarization and machine translation. It works by comparing a summary or translation generated automatically with a set of reference summaries. ROUGE score is summary accuracy evaluation metric which compares the human generated summary and machine generated summary. Following are the evaluations done by the proposed system.

Recall (R): Number of overlapping words / total words in reference summary
Precision (P): Number of overlapping words / total words in system summary.
F1_score (F1): (2* Precisions * Recall) / (Precision + Recall)

4.2 Result

The proposed model used BBC news dataset from Kaggle and image input containing the android information as an input. Fig .2 and Fig. 3 show the generated summary for the given input. .Fig.2 shows the output generated by the model when image is given as a input. Fig.3 shows the output generated when text is given as input.



INPUT:

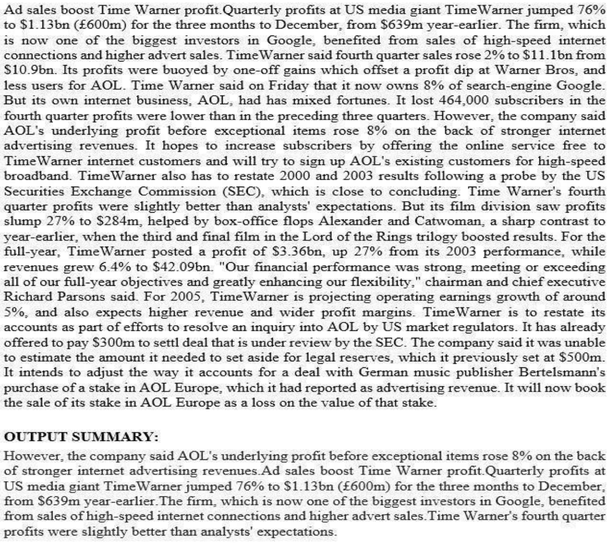
Android is the world's most popular mobile platform. With Android you can use all the Google apps you know and love, plus there are more than 600,000 apps and games available on Google Play to keep you entertained, alongside millions of songs and books, and thousands of movies.

Android devices are already smart, and will only get smarter, with new features you won't find on any other platform, letting you focus on what's important and putting you in control of your mobile experience.

OUTPUT SUMMARY:

Android devices are already smart, and will only get smarter, with new features you won't find on any other platform, letting you focus on what's important and putting you in control of your mobile experience. Android is the world's most popular mobile platform

Fig. 2 Output for image input



Ad sales boost Time Warner profit.Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier. The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL. Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the online service free to TimeWarner internet customers and will try to sign up AOL's existing customers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the US Securities Exchange Commission (SEC), which is close to concluding. Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins. TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to settl deal that is under review by the SEC. The company said it was unable to estimate the amount it needed to set aside for legal reserves, which it previously set at $500m. It intends to adjust the way it accounts for a deal with German music publisher Bertelsmann's purchase of a stake in AOL Europe, which it had reported as advertising revenue. It will now book the sale of its stake in AOL Europe as a loss on the value of that stake.

OUTPUT SUMMARY:

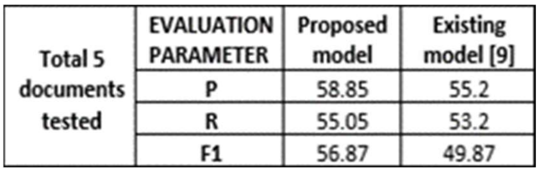
However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues.Ad sales boost Time Warner profit.Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier.The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales.Time Warner's fourth quarter profits were slightly better than analysts' expectations.

Fig. 3. Output for text input

4.3 Evaluation with ROUGE

For a single document, the proposed method for extractive text summarization is used. For the summarization of input data, the Textrank algorithm is employed. For analyzing generated summaries, ROUGE (Recall Oriented Understudy for Gisting Evaluation) is used. Precision, recall, and the F measure are the performance evaluation metrics used to evaluate ROUGE score.

Table 1 Performance comparison between results obtained by the proposed model and existing model



	EVALUATION PARAMETER	Proposed model	Existing model [9]
Total 5 documents tested	P	58.85	55.2
	R	55.05	53.2
	F1	56.87	49.87

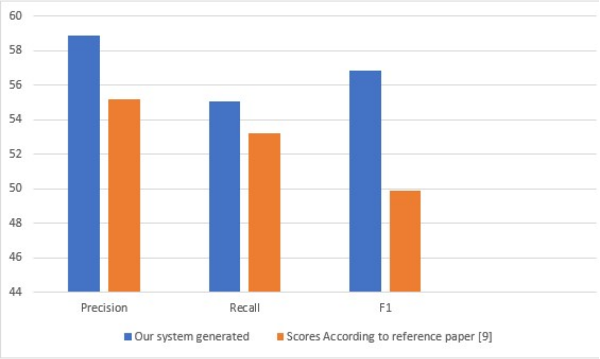


Fig. 5 Graphical comparison between results obtained by proposed model and previous model

Table 1 shows the comparison between the obtained results from proposed model and existing model. Fig.5 graphically shows how the proposed model is working better than the previous model. For total 5 documents the system performance is tested. After comparing the results of proposed model and existing models we came to the conclusion that our model is providing better precision, recall and f1 score than the existing model.

5. Conclusion and Further Work

The textrank algorithm generates a summary for the BBC news dataset combining the text and image summarization model and producing an integrated approach with the increased accuracy of 57.75% when compared with the traditional methods of summarization. The system generating a summary for both text and image input type combined were not existing. Also, the proposed model provides high ROUGH score even after combing the two model together.

Future work:
The proposed model already processes the summarization for text and image input type. The current system can be enhanced by integrating more classification techniques along with adding various functionalities. The proposed model could also be used to test on various different datasets to increase the versatility of the entire model.

References

[1] Li, W., & Zhuge, " Abstractive Multi- Document Summarization based on Semantic Link Network." IEEE Transactions on Knowledge and Data Engineering, (2021)

[2] Pisat, T., Bartakke, M., & Patil, H. "Synonym Suggestion System using Word Embeddings" 4th International Conference on Trends in Electronics and Informatics, (2020)

[3] Singh, P., Chhikara, P., & Singh, J, " An Ensemble Approach for Extractive Text Summarization." International Conference on Emerging Trends in Information Technology and Engineering (2020)

[4] Szucs, G., & Huszti, D. "Seq2seq Deep Learning Method for Summary Generation by LSTM with Two- way Encoder and Beam Search Decoder." IEEE 17th International Symposium on Intelligent Systems and Informatics (2019)

[5] Rahul, Adhikar, S., & Monika. "NLP based Machine Learning Approaches for Text Summarization." Fourth International Conference on Computing Methodologies and Communication (2020)

[6] Aciar, S. V., & Ochs, M. "Classifying User Experience based on the Intention to Communicate." IEEE Latin America. (2020)

[7] Alengadan, B. B., & Khan, S. S. "Modified aspect/feature-based opinion mining for a product ranking system." IEEE IT transactions (2018)

[8] Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. "Non- Query-Based Pattern Mining and Sentiment Analysis for Massive Microblogging Online Texts." IEEE Access. (2020)

[9] J.N.Madhuri, Ganesh Kumar.R, "Extractive Text Summarization Using Sentence Ranking" IEEE Access,(2019)

[10] Vrublevskyi, V., & Marchenko, O. "Grammar Error Correcting by the Means of CFG Parser." IEEE International Conference on Advanced Trends in information theory. (2019)