International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Rule-Based Extractive Summarization and Title Generation

Insiya AbdulSalam, Shirien K A, Surekha Mariam Varghese

Abstract*: Text summarying is a process by which the most important information from the source document is precisely found. It stands for the information condensed to a longer text. Text summary is broken down into two approaches: extractive summary and abstractive summary. The proposed method creates an extractive summary of a given text and generate an appropriate title for the generated summary. Extractive summary is generated through sentence selection by using Rule-based concept. Eight different features are considered to rank each sentence according to its importance. Ranking assigns a numerical measure to each sentence. After ranking, sentences that has high rank compared to others will be selected to form the summary. The frequently occurring bi-gram is selected as the title for the summary. The system performs better than existing extractive summarization techniques like Graph-based system and achieved a precision of 0.7

Keywords*: Extractive summarization, Abstractive summarization, Rule-based, Graph-based.

I. INTRODUCTION

Text summarisation is a Natural Language Processing (NLP) application field. In reality it is the condensed details of a longer text. The summary should contain the document's most important information, but its size should be lower than the document's. With the aid of this text summarization tool, a long news content can be condensed to a shorter version including all of the relevant information. In essence, there are two types of summary process: Extractive summary and Abstractive summary. Extractive summarying is a process where important phrases are copied as they are from the original document and these phrases represent the whole document. On the other hand, abstractive text description is a method in which a whole new set of sentences are generated that are different from the original document to convey the input document's general information.

The main purpose of text summarization is to present in a shorter version the most important concepts of the original document while retaining its key content, thereby enables the consumer to grasp large volumes of materials easily. Text review addresses the problem of choosing and generating coherent summaries of the most important sections of the text. This is remarkably different from the description produced by humans because humans can capture and compare profound meanings and the subject matter of text documents and replicate text alone. Automating a method like that is very hard to implement. The summaries give a glimpse of interesting information, which produces a smaller version of each document in the package. Reading description helps or does not determine whether to read the entire text, so it acts as time saver.

Different researchers have suggested various techniques for automated text description that can be categorized in two ways: extraction and abstraction. Extraction review is taking sentences or phrases with highest score from the given document, and put it up together in a new smaller version without modifying the source text. Process of abstraction description use various semantic and linguistic approaches to analyze and define wording. Majority of current automated text summary is based on extractive methods. Input automatic summarying is a complex piece of knowledge extraction process. When summarizing text based on extraction, the interpretation of different sentences from the source document is an important part. A Legal Summarizer that offers a description of the text with better coverage of the details was suggested. Much of the effectiveness of the document summing method is that it reduces the human effort to create the document description. To relate the deep meanings and understanding themes of large documents is a difficult process. Thus automation of such a skill overcome all difficulties and helps to interpret large documents.

II. LITERATURE SURVEY

Automatic summarying is a technique of reducing the size of a record for text for the program, producing an overview to the main initial paper points [1]. Methodologies that can generate a consistent description consider the factors, such as frequency, style writing, and syntax. Machine learning and data mining are concerned with the automated data summarization. The main agenda behind summarying is to search for a subset of data containing the entire set of information. There are basically two forms of description extractive functions [5] based on what the review process is centered on. The first is a general overview, aimed at providing a general description or abstract of the collection (whether papers, or picture sets, or images, news stories, etc.).

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Insiya Abdulsalam*, Computer Science Department, Mar Athanasius College of Engineering, Kothamangalam, India. Email: inusaam7@gmail.com
Shirien K A, Computer Science Department, Mar Athanasius College of Engineering, Kothamangalam, India. Email:shirienashraf2211@gmail.com
Dr. Surekha Mariam Varghese, Computer Science Department, Mar Athanasius College of Engineering, Kothamangalam, India. Email:surekh.var@gmail.com

The latter is a description appropriate for queries, also called domain-based summary, which resumes texts that are unique to a query. Depending on what the actual need is, summary systems will produce both the text summaries relevant for queries and the generic summaries created by the computer.

Vishal et.al 2019 suggested a phrase-level bag of terms with the normal weighted term frequency and inverse phrase frequency model, where phrase frequency is the number of phrases in the text in question where the term is present. Instead, these sentence vectors are graded by a similarity of questions and the highest scoring sentences are chosen as part of the explanation. Here the paradigm of information retrieval is specifically related to the model of summarization. Summarization is specific to queries. Non-stopwords, which appear most frequently in the text, can be taken as words of inquiry to produce a generic description. Because these words represent the document theme they generate a generic summary. Term frequency for sentence ordinarily 0 or 1 because typically same word for material does not show up in a given sentence several times.

Suanmali et.al.,2011 [9] proposed an automated text summary approach with the extraction of sentences using bubbling logic, genetic algorithm, semantine function labeling and their combinations to produce summaries of high quality. This research searched the usefulness of the genetic algorithm in the problem of optimizing selection of the features during training phase, and changes feature weights during the test phase. Fuzzy IF-THEN rules[3] were used for Weights alternating between important and insignificant features. Traditional methods[4] of extraction cannot define semantic relationships between concepts in a document. Thus, this research explored application of semantine functions marking identify and incorporate the semantic content in sentences into the summary method.

Canasai et.al., 2003 [11] suggested a practical approach to the creation of a description of the many important phrases from the original document. The purpose of this approach is to manipulate the properties of sentences, both local and global. Within a sentence, a nearby property could be seen as clusters of meaningful words, while the global property can be considered as relationships of all the sentences in the document. Both of these properties combine to obtain a single measure that reflects sentence informativity.

In 2002 Liu et al introduced the idea to use Latent Semantic Analysis in ATS. Taking the ideas from latent semantic indexing, they used the outstanding analysis of the value to the domain of text summation. The decomposition of single values is a very useful mathematical method used to find main multidimensional orthogonal data dimensions. It has implementations in many different fields and is known by various names: Karhunen-Loeve Transforming image processing, PCA processing the signal, and Latent Semantic Analysis (LSA).

III. PROPOSED METHOD

Automatic engraving of text is a challenging task in the area of information recovery. In summarizing text based on extraction, an important part is finding specific sentences from the source document. A Rule-Based Summarizer was proposed which provides a description of a text with increased coverage of the material. Every sentence of the document is defined by sentence score in this system. All sentences of the text are then evaluated according to their scores in a descending order. We extracted required number of phrases based on a compression rate of 30%. It has been shown that extracting 30% of phrases from original document can be as insightful as the complete text of a document. The summary phrasing is ultimately placed originally ordered.

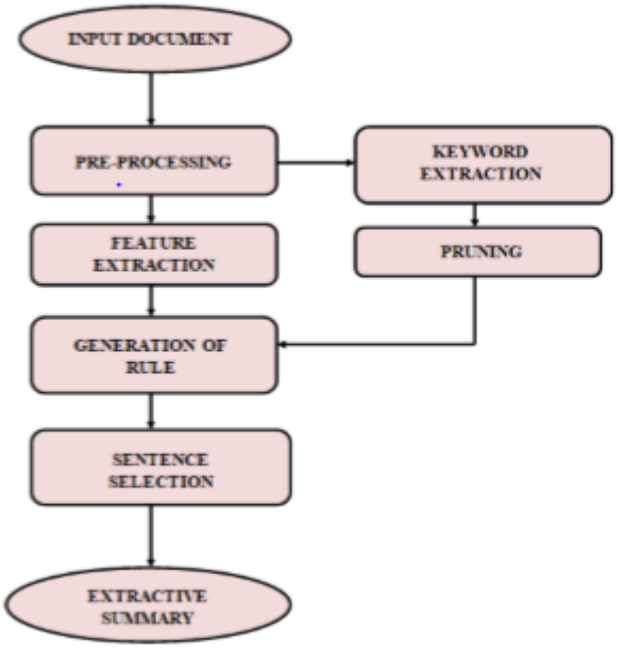

Fig. 1. Text Summarization Architecture

A. System Architecture

The text would be subject to different methods of extraction of features in this summary methodology, where those words are interpreted as the characteristic variable. A rule is written, and all sentences are transmitted through that code. Lastly, the phrases are sorted and only the top phrases are chosen based on the severity of the summary specified for summarise and title produced by considering bi-grams.

B. The Infrastructure

The framework created for the review comprises mainly the following measures

1) Data set and Preprocessing: About 20 documents to create automatic summarization of single documents. Every document contains approximately 8 to 60 sentences, averaging 28 sentences. We intended to generate an output, and the length of that description depends on the document input text. Record entering is actually of plain text format.

At this stage major four tasks are carried out: Sentence Segmentation, Tokenization, Removal of Stop Word and Word Stemming. Segmentation of sentences is the detection of boundaries and the detachment of source in sentences. Tokenisation separates the document join in single words. Next, Removing Stop Words, stop words are words that often appear in a document but have less significance in identifying important content of the document such as ' a, ' ' an, ' ' the, ' etc.

Word Stemming is the final step of pre-processing; word stemming is the method of eliminating the prefixes and suffixes of each word.

2) Sentence Length: This function is useful in filtering out short sentences including datelines and names of writers which are generally find in the press. Belonging to the summary, the short phrases are not conventional. We use the term length, which is the percentage of word count appearing in the phrase over the word count appearing in the document's longest sentence.

3) Term Weight: The word frequency within a text is often used to measure the meaning of the sentence. A sentence's score could be computed as the number of word score in the sentence.

$$w_i = tf_i * isf_i = tf_i * \log \frac{N}{n_i}$$

(1)where tf_i is the term frequency of word i in the text, N is the total number of sentences, and ni is number of sentences in which word i appears.

4) Sentence to Sentence similarity: This function determines parallels between sentences . The relation between S and one other sentence is determined by the cosine similarity measure with a corresponding value range 0 and 1 for each sentence S. The term weight wi and wj of term t to n term in sentence Si and Sj are represented as the vectors. The similarity of each pair of sentences is calculated based on the formula for similarity. The score of this function for a sentence S is obtained by measuring the ratio of the summary of sentence similarity of sentence S over the summary limit to each other sentence.

5) Proper Noun: Sentence that includes more correct nouns (name entity) is significant, and is most likely contained in the description of the text. The score for this function is determined as the ratio over the length of the sentence to amount of proper nouns in sentence.

6) Thematic Word: Amount of thematic word in paragraph, this attribute is a key, since words that appear frequently in a text are likely related to subject matter. The number of thematic words shows the terms with the highest relativity possible. As a thematic we used the top 10 most frequently used word material. The score for this function is calculated as the number-ratio of thematic words in the phrase to total feasible description of thematic words in the phrase.

7) Numerical Data: Number of pieces of data in sentence, phrase consisting of number figures are relevant and is most likely included in the record of the document. The score for this function is calculated as a ratio of the amount of numerical data occurring in sentence over the length of the sentence.

8) Sentence Feature: Sentence's score is based on several sentence characteristics, and helps to summarize the documents by sorting sentences based on these scores.

C. Title Generation

After filtering the sentences, several sentences are left over using the configurations mentioned in previous subsections. Terms that appear in the computer-generated title are taken from those sentences by considering the most significant bigrams that have occurred.

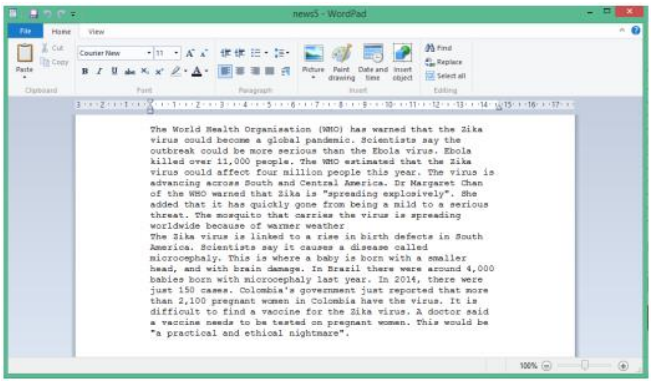

Fig. 2. Text Summarizer Input

IV. EXPERIMENTAL RESULT

Experiments done on Intel core i5 processor, 64-bit OS, 4GB RAM. The Java is used as the programming language. The NetBeans IDE 8.0.2. Is used as a medium for the execution of experiments. This applies the extractive description methodology and the GSM technique. Comparison of summaries generated by the extractive summarization technique with summaries generated by GSM technique. The methods used to determine the efficiency of the rule-based extractive summary program are re-call, precision and f-measure. Recall is measured by taking the proportion of appropriate sentences in summary. Precision tests the proportion of sentences correct in rule based extractive summary. The F-measurement is calculated by taking weighted harmonic recall mean and accuracy values .Recall, Precision and F-Measure were determined using GSM Summarizer and Extractive Summarizer based on the Rules for fifteen set of documents from the DUC data store.

GUI is a form of connection that allows users to use graphical icons and visual indicators to communicate with electronic devices. In a GUI, activities are usually carried out by direct manipulation of the graphical elements. Here the project uses PYQT for creating the user interface.The system is implemented using Python. Algorithm is purely based on Python's Natural Language Tool Kit (NLTK).We choose NLTK is by considering several advantages over the older natural language processing tools is basically the efficiency. It was created to support education. Most advanced techniques like Tokenization, POS tagging etc are included in NLTK. Another advantage of NLTK as a programmer is considered is the efficient scripting employed in NLTK.

V. PERFORMANCE ANALYSIS

Performance Analysis deals with the measurement of response time and through-put of our software and comparing it with existing summarization techniques with the help summarization benchmarking tools. Our summarization software relate the specific sentences, reduce the sentence lengths and thereby reducing the whole document into well summarized document.

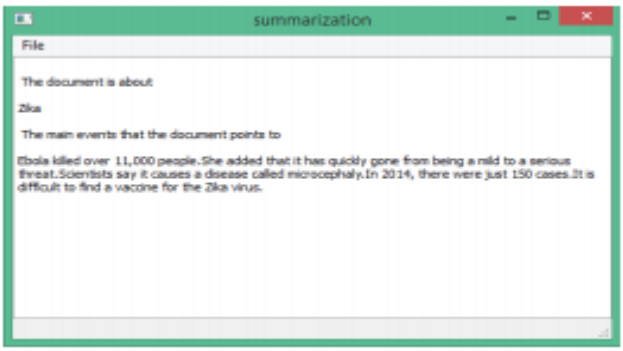

Fig.3. Text Summarizer Output

Apart from the Automation techniques, human skills can go deep into the summarization topic and relate the sentences in more logical tokens which can be constrained to obtain a summary which is more user convenient. Researches related to this technique are still going on and the new features are taken into account for improving the deepness in relating sentence and obtaining well formed, user convenient, more compact summary. This applies the rule-based extractive summary approach and the current GSM method. Summaries generated by the extractive method are compared with GSM generated summaries. The main methods of system performance assessment are recall, accuracy, and f-measurement. Recall is determined by calculating the proportion of corresponding summary sentences. In summary, accuracy is calculated as the proportion of correct sentences. Applying GSM Summarizer and the Rule Extractive Summarizer, F-measurement determined by weighted harmonics recall mean and accuracy values. Calculated re-call, F-measure and accuracy for fifteen DUC dataset sets of documents.

Precision (also referred to as positive predictive value) is the fraction of significant instances among the retrieved instances, while recall (also referred to as sensitivity) is a fraction of the actual instances over the total number of important instances. Therefore, both precision and recall are based on an understanding and a measure of relevance. The equations used to calculate recall and precision were:

$$Precision = Correct/(Correct + Wrong) \quad (2)$$

$$Recall = Recall / (Correct + Missed) \quad (3)$$

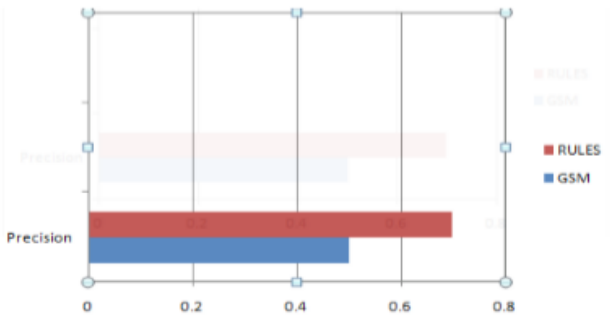

Fig. 4. Precision

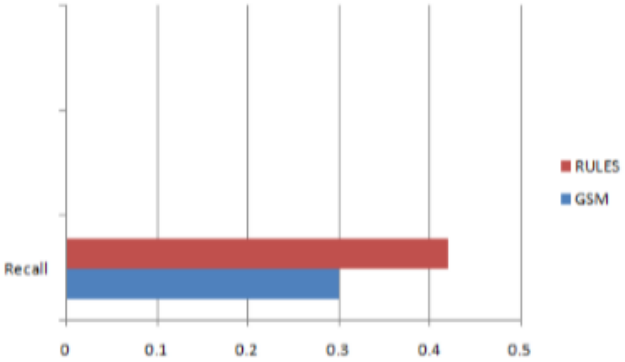

Fig. 5. Recall

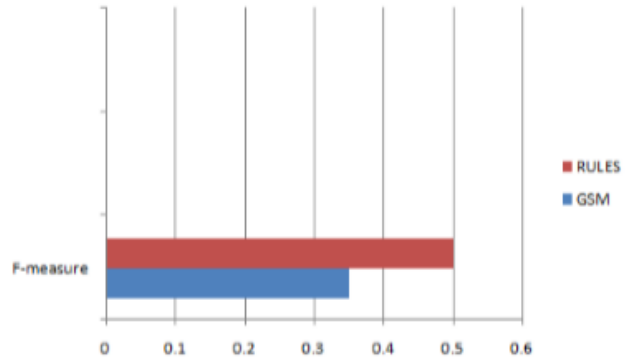

Fig. 6. F-measure

VI. CONCLUSION

The deployment of summarizing software brings an advanced means of obtaining a well compact, perfectly defined, tightly constrained summary about the topic of the document along with efficient title. It will reduces all the human efforts for summarization of a document for those who want to know an outline of a large document. We have proposed a Rule-Based Extractive Summarizer which provides an overall idea of a document with good coverage of information. Twenty news articles from BBC data set were the Inputs to Extractive summarizer. Each document sentence is in the form of feature vector attributes. Results produced by extractive summarizer were compared with the summarizers available. It was found that the extractive summarizer provides better average values of precision, re-call and f-measures than those of current summarizers.

REFERENCES

1. Jasmeen Kaur and Vishal Gupta, “Effective Approaches for Extraction of Keywords”. International Journal of Computer Science Issues (IJCSI), vol. 7, issue 6, November 2010.
2. Generic Text Summarization Using Local and Global Properties of Sentences, Canasai Kruengkrai and Chuleerat Jaruskulchai Proceedings of the IEEE/WIC International Conference on Web Intelligence 2003 IEEE
3. Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan “Fuzzy Genetic Semantic Based Text Summarization”. 2011 Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing.

4. Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques". Journal Of Emerging Technologies In Web Intelligence, vol. 2, no. 3, August 2010.
5. Ladda Suanmali, Mohammed Salem, Binwahlan and Naomie Salim, "Sentence Features Fusion for Text Summarization using Fuzzy Logic". Ninth International Conference on Hybrid Intelligent Systems, IEEE, 142-145, 2009.
6. Khosrow Kaikhah, "Text Summarization using Neural Networks". Proceedings of Second International Conference on Intelligent Systems, IEEE, 40-44, Texas, USA, June 2004.
7. F.Canan Pembe and Tunga Gungor, "Automated Query-Biased and Structure-Preserving Text Summarization on Web Documents". Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, June 2007.
8. Canasai Kruengkari and Chuleerat Jaruskulchai, "Generic Text Summarization using Local and Global properties of Sentences".Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03), 2003.
9. Dharmendra Hingu, Deep Shah and Sandeep S. Udmale, "Automatic Text Summarization of Wikipedia Articles". International Conference on Communication, Information and Computing Technology (ICCICT), IEEE, 2015.
10. Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan, "Feature-Based Sentence Extraction using Fuzzy Inference Rules". International Conference on Signal Processing Systems, IEEE, 2009.

AUTHORS PROFILE

Insiya AbdulSalam received Bachelor of Technology in Computer Science and Engineering from Government Engineering College Palakkad, in 2017 and currently pursuing Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. Her research interest is in Deep Learning and Data Mining.

Shirien K A received a Bachelor of Technology in Computer Science and Engineering from KMEA Engineering College, Edathala in 2018 and is currently pursuing a Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated with APJ Abdul Kalam Technological University. Her research interest is in Deep Learning and Data Mining.

Dr. Surekha Mariam Varghese currently heads the Computer Science and Engineering department, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. n 1990, she received her B-Tech Degree in Computer Science and Engineering from the College of Engineering, Trivandrum affiliated with the University of Kerala, and M-Tech in Computer and Information Science from the University of Cochin, Kochi, in 1996. In 2009, she earned a Ph.D. in Computer Security from the University of Science and Technology in Cochin, Kochi. She has some 27 years of experience in teaching and research at various institutions in India. Her research interests include machine learning, network security, database management, algorithms and data structures, operating systems, and distributed computing.