MULTI DOCUMENT AUTOMATIC EXTRACIVE TEXT SUMMARIZATION USING N-GRAM APPROACH

S.Sudha Lakshmi
Research scholar, Dept. of Computer Science
SPMVV, Tirupati, India.
s_sudhamca@yahoo.com

Dr. M. Usha Rani
Professor, Dept. of Computer Science
SPMVV, Tirupati, India.
musha_rohan@yahoo.com

ABSTRACT- Multi-Document Summarization represents a collection of documents with a concise version of text by capturing the relevant information and filtering out the redundant information. Multi-Document Summarization consists of two prominent approaches such as extractive and abstractive summarization. Extractive summarization systems aims to extract salient sentence , keywords or passages from original source text, while abstractive summarization systems aim to concisely paraphrase the content of the documents or new words which are not from original source text.

A simple set selecting key words from original text and generating summary for multi documents is a huge task in Extractive summarization process. This process internally connected with 4 tasks. First is to pre-process the information using stop word removal and tokenization. In second phase information verification i.e., verifying word count in each under start and stop word model. In third phase with the help of word count, the sentence selection is implemented. Finally, sentences selected were clustered and an extractive summary was generated. This approach of word count based summary generation was made possible using N-GRAM model approach relies on word count and frequency of words in sentences. In this process a chain-based operation was performed for word connectivity. The performance of proposed model Evaluated on DUC 2002, DUC 2003 and DUC 2004 for multi documents using ROUGE-1, ROUGE- 2, ROUGE-3 and ROUGE-L performance metrics.

KEY WORDS : *Keywords, summary, multi-documents, extractive summarization, N-GRAM, chain.*

1. INTRODUCTION

Along with the growth of the internet and big data, making people overwhelmed by the large information and documents on the internet. This triggers the desire of many researchers to develop a technological approach that can summarize texts automatically. Automatic text summarization generates summaries containing important sentences and includes all important relevant information from the original document [1-2]. So the information quickly arrives and does not lose the original intent of the document [3]. The area of text summarization research has been studied since the mid-20th century, which was first discussed openly with a statistical technique namely word frequency diagrams. Many different approaches have been created to date. Based on the number of the document, there are single and multi-document summarization techniques. Based on the output of the summary there are the extractive and abstractive results.

A single document produces a summary that is sourced from one source document and the content described is around the same topic. While the multi-document summarization is taken from various sources or documents that discuss the same topic [4-7] made text summarizing in a single document using TF-IDF and [8] designed automatic text summarizing in a single document using the Main Concepts. [9] summarized multiple documents by the pattern-based summarization (Patsum) method on the 2004 DUC dataset and showed that the results outperformed not only the term-based method but also the ontology-based method [10]. summarized multiple documents using Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NMF) and the results outperformed the state of the art in precision and recall. [11] proposes the summarization of the Arabic single document which produces a fairly informative extractive summary combining machine learning and score-based approaches that evaluate each sentence based on a combination of semantics and statistics. The results are superior in terms of precision metrics, memory, and F-scores, the disadvantage is that it has not optimized the weight of the features. [12] minimized redundancy in multi-document summarizing by the Shark Smell Optimization (SSO) method and the performance results were far better than the previous summary method.

Extractive summarization shortens the contents of original text which results in important sentences or words . The usual problem raised from the extractive summarization research at first was determining the position of the sentence and the frequency of words in the text. The next experiment raised the extraction problem which is known as the Information Extraction (IE) technique to produce a summary with more specific results to increase accuracy. One example of an automatic summarization system that has been developed by adopting IE techniques is RIPTIDES, which functions to summarize news based on scenarios chosen by the user. Research by using a rule base produces the best average precision, f-measure, and recall values for Rule-Based Summarizers but has not yet been tried on broader data. Furthermore, there are extractive summarizing studies using neural networks, which in recent years have achieved greater popularity than conventional approaches, some of these studies. Research conducted by used a deep learning technique namely Feed Forward Neural Network (FFNN) to summarize a single document in a legal document that has the advantage of producing an extractive summary without the need to create features or domain knowledge and perform well as measured by the Rouge score and produces a coherent summary, but weak in terms of simplifying complex and long sentences.

In contrast to extractive summarization, sentences generated by abstractive summaries are new sentences or commonly called paraphrases which produce summaries using words that are not in the text. Abstractive summaries are very complex and relatively more difficult than extractive summaries because producing abstractive summaries requires extensive natural language processing [13] Approach techniques in abstractive summary are generally grouped into two categories namely the linguistic approach and the semantic approach. Examples of methods that use linguistic approaches such as information-based methods and tree-based methods. While examples of methods that use semantic approaches such as template-based methods and ontology-based methods. More recently research on abstractive summarizing has been inspired by the encoder-decoder framework. Besides being believed that this model is smoother, the encoder-decoder framework is also convenient in adjusting parameters automatically.

2. LITERATURE SURVEY

The Multi-document text summarization is a complex task that consists of various sub-tasks. Each of the sub-tasks directly affects the ability to generate high quality summaries. In extraction based summarization the important part of the process is the identification of important relevant sentences of text. Use of fuzzy logic and Machine learning as a summarization sub-task improved the quality of summary by a great amount. An example of a method that uses a fuzzy-based approach with classic Zadeh's calculus of linguistically quantified propositions which addresses trend extraction and real-time problems where the results are superior in t-norm evaluation, but weak in semantic problems because the semantic results of other t-norms are unclear and unclear can be understood. Fuzzy Formal Concept Analysis [8] (Fuzzy FCA) which addresses semantic and real time problems where the results excel at evaluations in f-measures with optimal recall and comparable precision. An example of a method that uses a machine learning approach is Incremental Short Text Summarization (IncreSTS) by which has better outlier handling, high efficiency, and scalability on target problems. Rank-biased precision-summarization (RBP-SUM) by which has advantages in overcoming redundancy by evaluating using rouge, but this method can only produce extractive summaries.

Text summarization is a formidable challenge in the field of Natural Language Processing (NLP) because it requires precise text analysis such as semantic analysis and lexical analysis to produce a good summary. A good summary, in addition, must contain important information and must be concise but also must consider aspects such as non-redundancy, relevance, coverage, coherence, and readability. Where to get all these aspects in a summary is a great challenge.

There are a numerous approaches for extractive multi-document summarization such as Graph, Cluster, Term-Frequency, Latent Semantic Analysis (LSA) based etc. Now research has shifted towards abstractive summarization and real-time summarization. But the summaries generated are not much relevant to original text which leads to inaccuracies. This is because abstractive summaries are more complex and complicated than extractive summaries. So extractive summaries are easier to give expected and better results than abstractive summaries. However, extractive summarization is also still in great demand as evident extractive research still exists in the last two years. This indicates the possibility that there are still opportunities or loopholes to improve.

A clear literature study is demanded as a means for the advancement of research in the field of text summarization. Literature study conveys a brief discussion of types ,approaches, techniques in the field of automatic text summarization. Types based on input ie single or multi documents ,based on output of the summary as extractive or abstractive grouped approaches to statistics, machine learning, semantic-based, and swarm intelligence. Another survey was conducted which is about summarizing extractive texts that focus on uncovered techniques, presents a list of strengths and weaknesses in a comparison table, alluding to a little about evaluations and future trends.

Some other review articles only cover smaller sections, for example only about approach techniques, the methods for evaluations , or discuss the topic of extractive or abstractive text summarization. So researchers, especially those who are new in studying this field, need to work hard and may have difficulty doing a thorough review. Therefore the objectives of this study are as follows: a) to identify and analyse research topics/trends in the field of summarizing texts and to classify them b) to provide an overview of the various approaches to summarizing texts (where the strengths and limitations of the commonly used approach are also highlighted c) to briefly explain the methods that already exist in this field, both frequently used and the latest methods d) to explain the pre-processing stages that already exist and what features have been used e) to discuss what problems have been the challenges in the field of text summarization so far and what problems have been solved or have not been resolved properly f) to briefly discuss existing evaluation techniques in summarizing text, as well as the data sets that have been used g) to present recommendations for future development of text summarization research.

3. METHODOLOGY

Multi document text summarization(MDS) for extractive summary consists of the selection of important sentences is the main step. For this purpose, the N-gram approach was used to generate an extractive summary which is considered as a label for further process. An N-gram can be termed as a series of N-words such as Uni-grams, Bi-grams, etc. Later the performance of the proposed model is evaluated using ROGUE scores which generously accessed N-grams. The N-Gram approach proposed in this paper requires a probabilistic reference model for sentence selection, which considers weight and length as parameters. To estimate each word's probability and generate word pairs (previous word, current word), it uses frequency weight component calculation, and then verifies whether the sentence should be part of summary or not, afterwards the extractive summary is generated.

Algorithm 1: Extractive Summarization Model based on N-grams model

Input: Merged Multi documents, set sentence length, set Start and Stop tokens for each sentence

```
STEP 1:for all sentences(S) do

STEP 2: Get the length of each sentence using a total count of words in the sentence.

STEP 3: Verify the Sentence length and calculate the Frequency Weight Component of words in each sentence

STEP 4: if the Sentence length parameter is matched or greater than the static word count of each sentence
```

STEP 5: Select the sentence

STEP 6: else

STEP 7: Update sentence count go to Step 2

STEP 8: end if

STEP 9: end for

Output: Extractive Summary.

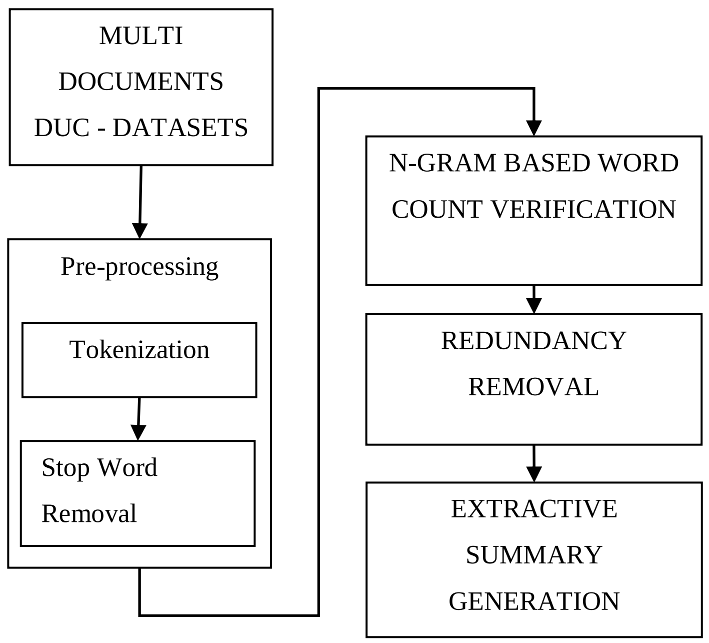
MULTI DOCUMENTS DUC - DATASETS
Pre-processing
Tokenization
Stop Word Removal
N-GRAM BASED WORD COUNT VERIFICATION
REDUNDANCY REMOVAL
EXTRACTIVE SUMMARY GENERATION

Fig 1: N-GRAM Model for MDS

This paper deals with a little bit of probability. In order to build vocabulary using deep learning models, a huge corpus of trained data is required. Therefore, building an NLP model by predicting the words in a sentence with the probability of a word in a series of words plays a significant role.

Based on the mathematical problem formulation Total merged text (TM_{text}) is given as an input to the system, then tokenization, contraction mapping, and stop word removal were applied as pre-processing approaches on the original merged N-Gram model-based summarization was performed on the merged pre-processed text. Sentence length is considered as a parameter for N-gram approach.

An N-gram is used for sentence selection, and Frequency Weight Component (FWC) was calculated to perform summarization. Therefore, the selection of sentences is given as

$$P_{text}(S) = \prod_{FWC} P(word) \qquad (1)$$

Where $P(word) \in TM_{text}$

Each paragraph eliminates the stop words and outcomes in $P_{text}(S)$ as a pre-processed sentence of the document. The concatenation of P(word) results in new sentences, and for that, frequency weight components for each pre-processed word are calculated. But the data obtained from (1) results in continuous form as it extracted from the sentence, but logically it was discretized one and represented as

$$\log(P_{text}(S)) = \sum_{FWC} \log\big(P(word)\big) \qquad (2)$$

The pre-processed stage eliminates the unwanted words from the sentences, results in discretized one and was subjected to likelihood test to get an extractive summary and verifies in P_{text} whether it is present in the Total merged text (TM_{text}) or not using Average log-likelihood (avgll) is given by (3)

$$avgll = \sum_{FWC} \log\big(P(word)\big) \cap TM_{text} \qquad (3)$$

The existing approaches used similarity measures to identify the duplication of information from the input documents.

In this proposed hybrid approach, redundancy elimination was performed due to the limitation of the length in summary sentences based on the likelihood ratio of the text. Then, the extracted words from the original text and verified words obtained after redundancy elimination chain rule were applied to compute the joint probability of words in a sequence.

$$Prob(sentence) \\ = Prob(X1)Prob(X1|X2)Prob(X1|x2|X3)Prob(X1) \ (4)$$

In equation (4) probability of each word (Prob(X1)) and the probability of word pairs (Prob (X1|X2)) was calculated, and concatenation of word pairs was performed to generate the sentences. Here X1, X2... Xn are words in a sentence.

Then obtain the sentences with exact or more than the length of words required. Next eliminate the sentences that have not reached the maximum length of static word count for new summary generation (GS). Finally, the union of sentences was accessed mathematically for extractive summary generation using the selected sentences.

GS

$$= \bigcup_{\mathrm{FWC}}^{\mathrm{L=100}} \begin{bmatrix} \mathrm{Prob(sentence1);} \\ \mathrm{Prob(sentence2)} \ldots \mathrm{Prob(sentence)} \end{bmatrix} \qquad (5)$$

4. Results

Deep learning based RBM model [9], RBM-FUZZY [9] and Deep learning-based N-Gram approach as proposed model for extractive summarization. The performance metrics observed are R1(Rouge 1), R2(Rouge2), R3 (Rouge3)and RL(Rogue -L)

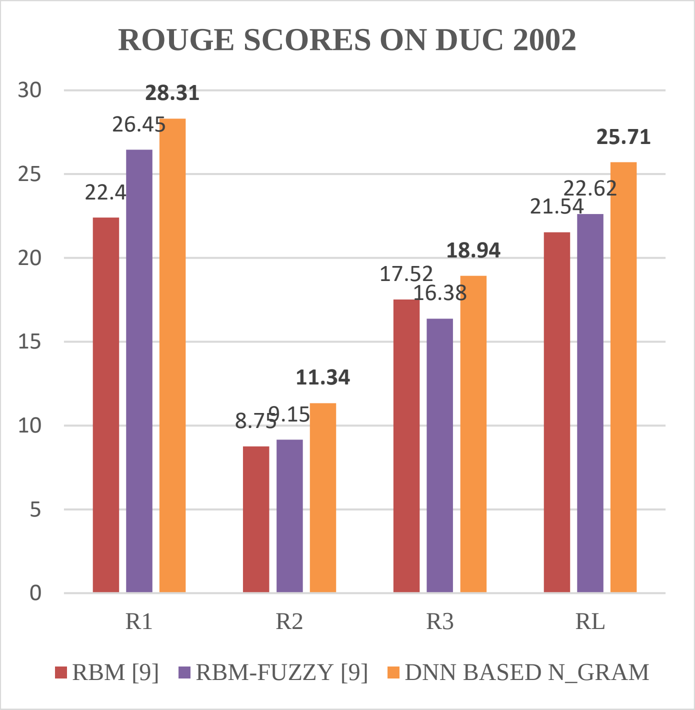

Fig 2:Performance metrics of various algorithms on DUC 02

Fig 2 shows that the proposed N-GRAM model for extractive summarization for multi documents attains highest score in ROUGE -1(Uni-grams)R1 28.31 when compared to existing models RBM and RBM-Fuzzy on benchmark Dataset DUC 2002.

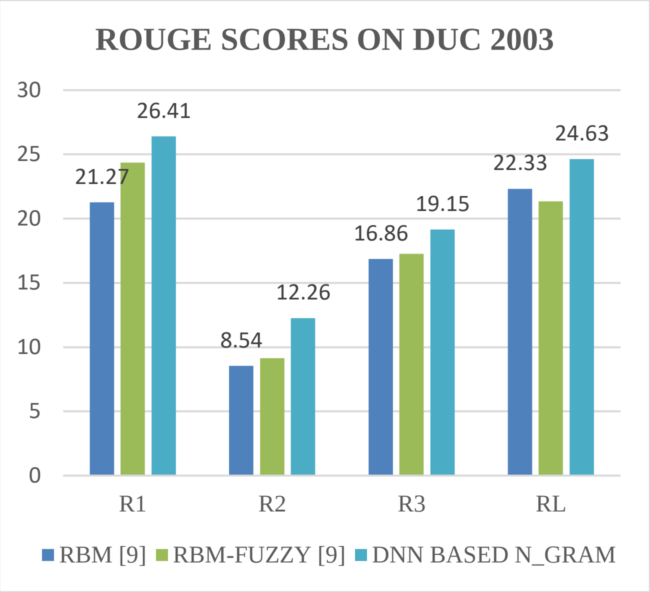

Fig3:Performance metrics of different algorithms on DUC03

Fig 3 shows that the proposed N-GRAM model for extractive summarization for multi documents attains highest score in ROUGE -1(Uni-grams)R1 26.41 when compared to existing models RBM and RBM-Fuzzy on DUC 2003.

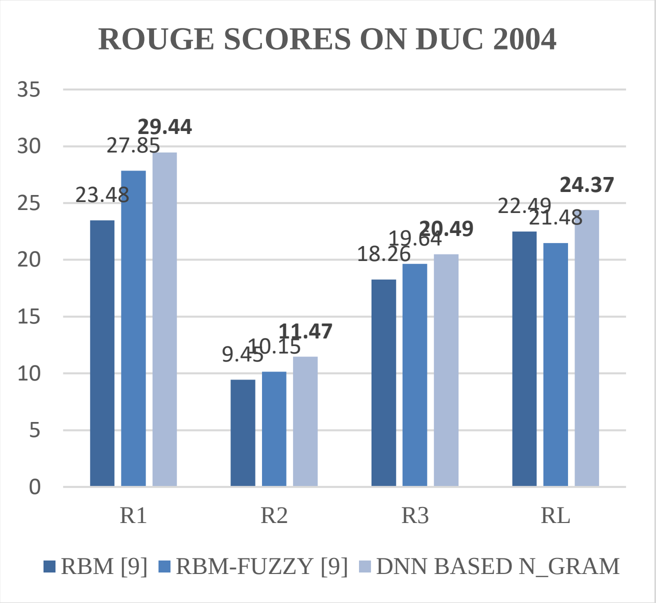

Fig4:Performance metrics of Various algorithms on DUC 04

Fig 4 shows that the proposed N-GRAM model for extractive summarization for multi documents attains highest score in ROUGE -1(Uni-grams)R1 29.44 when compared to existing models RBM and RBM-Fuzzy on DUC 2004.

5.CONCLUSION

In this paper, the proposed multi document text summarization minimizes the training time for extractive summarization and compared with existing approaches like RBM and RBM-Fuzzy. The proposed model was effective by 1.5-2.4 times than existing models in terms performance metrics Rouge scores R1(ROUGE 1),R2(ROUGE 2),R3 ROUGE 3,RL (ROUGE-L). The results shows that N-GRAM based extractive text summarization model not only generates efficient summary but also simplifies the complexity also. In future work, this model can extended to abstractive summaries or hybrid model to improve the quality in summary.

REFERENCES

[1] Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, Kochut K. Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268. 2017 Jul 7.

[2] Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey. Artificial Intelligence Review. 2017 Jan;47(1):1-66.

[3] Murad MA, Martin T. Similarity-based estimation for document summarization using Fuzzy sets. International Journal of Computer Science and Security. 2007 Nov;1(4):1-2.

[4] Qiang JP, Chen P, Ding W, Xie F, Wu X. Multi-document summarization using closed patterns. Knowledge-Based Systems. 2016 May 1;99:28-38.

[5] John A, Premjith PS, Wilscy M. Extractive multi-document summarization using population-based multicriteria optimization. Expert Systems with Applications. 2017 Nov 15;86:385-97.

[6] Widjanarko A, Kusumaningrum R, Surarso B. Multi document summarization for the Indonesian language based on latent dirichlet allocation and significance sentence. In2018 International Conference on Information and Communications Technology (ICOIACT) 2018 Mar 6 (pp. 520-524). IEEE.

[7] Khan A, Salim N. A review on abstractive summarization methods. Journal of Theoretical and Applied Information Technology. 2014 Jan 10;59(1):64-72.

[8] Naik SS, Gaonkar MN. Extractive text summarization by feature-based sentence extraction using rule-based concept. In2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) 2017 May 19 (pp. 1364-1368). IEEE.

[9] Lakshmi, S.Sudha and Rani, M. Usha, Multi-Document Text Summarization Using Deep Learning Algorithm with Fuzzy Logic (February 7, 2018). 2018 IADS International Conference on Computing, Communications & Data Engineering (CODE)

[10] R. Abbasighalehtaki, H. Khotanlou , M. Esmaeilpour, Fuzzy evolutionary cellular learning automata model for text summarization Swarm Evol. Comput., 1–16 (2016)

[11] K. Alsabahi, Z. Zuping, M. Nadher A hierarchical structured self-attentive model for extractive document summarization (HSSAS), IEEE Access XX (2018)

[12] M. Allahyari, S. Pouriyeh, M. Asse fi, S. Safaei, E.D. Trippe, J.B. Gutierrez, K . Kochut, Text summarization techniques: A brief survey Int. J. Adv. Comput. Sci. Appl., 8 (2017)

[13] S.A. Babar, P.D. Patil, Improving performance of text summarization, Procedia Comput. Sci., 46 (2015), pp. 354-363