Bulletin of Pure and Applied Sciences
Zoology (Animal Science), Vol.43B, No.2S,
July-December 2024: P. 443-448

Print version ISSN 0970-0765
Online version ISSN 2320-3188

Original Research Article
Content Available online at: www.bpasjournals.com

FEATURE BASED EXTRACTIVE SUMMARIZATION (FBES) TECHNIQUE FOR LONG TEXT DOCUMENTS

***B. Lavanya**
Associate Professor,
Department of Computer Science,
University of Madras,
Chennai, India.
lavanmu@gmail.com

U. Vageeswari
Research Scholar,
Department of Computer Science,
University of Madras,
Chennai, India.
uvageeswariphd@gmail.com

Abstract – Text summarization is a sensible area of research, owing to the increased utilization of text data. The process of presenting the content in a condensed form while maintaining the data's integrity is summarization. The development and expansion of automatic text summarization techniques are remarkable these days, due to their widespread applicability. The lengthy information must be processed such that the text summary is created. The text summarising research problem has been addressed in a large number of published studies, but there are not many solutions for processing lengthy documents. The pre-processing, the computation of inter-sentence similarity, and the generation of the summary are the three key phases that are discussed in this research article, which presents the FBES technique for the summarization of lengthy documents. The inter-sentence similarity is computed by the degree of information and ROUGE score of sentences. The generated summaries are evaluated against the gold summary and the results prove that the generated summary shows better ROUGE-N, cohesion, sensitivity and readability scores. The average best ROUGE-1 and ROUGE-2 scores of the proposed work are 44.71 and 18.56 respectively, these scores are roughly 3% higher than those received by the present method.

Keywords- Text processing, extractive summary, long documents, readability, ROUGE.

1. INTRODUCTION

Text summarization is crucial for a number of reasons, including the quick retrieval of vital information from a lengthy text, and the simple and speedy loading of the most essential information [1]. Recent advances in artificial text summarizing algorithms are remarkable due to their widespread use. This work helps summarize long text documents more efficiently than current methods. Search engines and news websites use text summary most [2]. One way to categorise text summarization separates the summarization process into extractive and abstractive categories [3]. Text summarization can be divided into a variety of categories depending on function, genre, summary context, sort of summarizer, and quantity of documents [4]Abstractive summarization rewrites the source text to create new terms. Abstractive summarization needs real-world information and semantic class analysis, making it harder than

extractive[4]. On the other hand, extractive summary involves minimal computational complexity with reasonable intelligibility. Hence, extractive summarization is preferable to abstractive summarization [5]. Both kinds of summarization techniques should ensure that the sentences of a summarized document should retain the theme and concept of the original document while avoiding text redundancies.

The remaining sections of the work are organized in the following way. Section 2 reviews the existing literature concerning text summarization and the proposed extractive text summarization technique is elaborated in section 3. The performance of this work is evaluated in section 4 and the paper is concluded in section 5.

2. REVIEW OF LITERATURE

This section studies the existing literature with respect to text summarization. In [6], a topic-aware text extractive and abstractive summarization approach based on Bidirectional Encoder Representation from Transformer (BERT). This work infers from topics and text summarization is carried out.A review of Arabic text summarization is presented in [7]. This work studies text summarization approaches with respect to deep learning models.A text summarization model along with information retrieval presented by employing deep learning is presented in [8]. This work is based on three phases such as information retrieval, template formation and text summarization. The textual data is retrieved by Bidirectional LSTM (BiLSTM) and embeds it in a semantic vector. Giga word and DUC corpora are employed for validating the work performance. In [9], a scheme employing multi-grained information is presented to improve text summarization. This work forms a fine-grained factual graph while maintaining relations with facts. An extractive text summarization model is proposed in [10], which is based on the Distant Supervision-based Machine Reading Comprehension (DSMRC) model. In [11], a survey of automatic text summarization concerning both abstractive and extractive approaches is discussed. The work proposed in [12] is based on enhanced attention-based bidirectional LSTM and presents a structure by sequence to sequence while improving the correlation between the summarized and the source text. The proposed work is presented in the following section.

3. Feature Based Extractive Summarization (FBES)

Extractive summaries from long documents are difficult because of the volume of words. This activity processes scholarly publications, averaging 5000–7000 words. The major objective is to improve context intelligibility, which is challenging in lengthy documents, especially for novice readers. Figure 1 shows the overall flow of the proposed summarization approach.

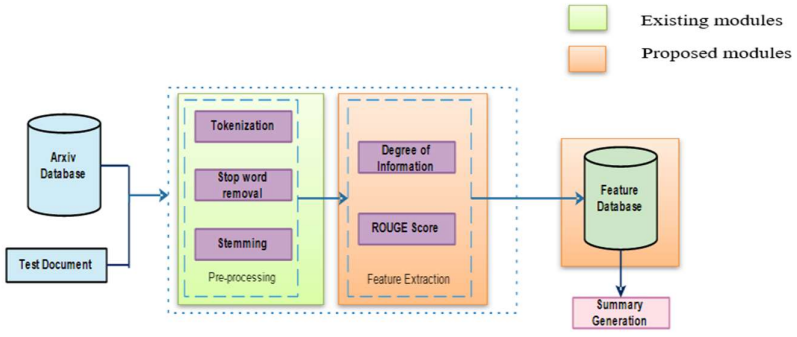

Fig.1. FBES technique for long documents

3.1 Long Document Summarization

This work focuses on long document summarization, which uses pre-processing, phrase similarity computation, and summary representation to construct an extractive summary.

3.1.1 Pre-processing

Data pre-processing begins with sentence extraction from the input document.

Academic articles have an introduction, literature survey, suggested study, performance analysis, and conclusion. The summary should include most of the major points in these sections.

$Doc_i = \{S_1, S_2, S_3, \dots S_m, S_n\}$ (1)

Here, n represents the total count of sentences in the document and m is the m^{th} sentence of the document i.

$TS = \{T_1, T_2, T_3, \dots, T_p\}$ (2)

In the above representation, p is the total count of terms available in the document and TS is the term set. The overall algorithm of the work is presented below.

Feature-Based Extractive Summarization (FBES) Algorithm

INPUT: *LTD- Long Text Documents, sn -Summary Length*
OUTPUT: *S- Summary*
BEGIN
//Pre-processing
Doc = []

For Each *doc in LTD*
    $Doc_i = Extract\ sentences\ (doc)$;
End for

For Each *sen in Doc*
    T=Tokenize sentence; // T-- Tokens
    STS=Remove stop words; //STS – Standard Token Set
    *Info[sen]=Compute_*$Degree\_info(sen)$
End for

For Each s_i *in Doc*
    For s_j *in Doc*
     *Sim[i,j]=Compute_*$similarity(s_i, s_j)$
End for

Sort (S) // Arrange Sim_{score} *in ascending order;*
For *i=0 to sn*
    Summary = S[i]
End for

Return Summary;
END

Sentence Extraction

Extract sentences (LTD) LTD—Long Text Document
    S=[]
    S =split (LTD)
Return Doc

Compute_Similarity

Compute_Similarity(S_i,S_j)

$$sim\_score = \frac{Compute_{Deg\quad info(s_i)}.\ Compute\_Degree\_info\ (s_j)}{\sqrt{Compute\_Degree\_info\ (s_i)^2 \cdot Compute\_Degree\_info\ (sj)^2}}$$

Rerurn sim_score

Compute_Degree_info

Compute_Degree_info (T)

$$Degree\_info = OF(T_i) \times \log\left(\frac{n}{n_k}\right)$$

Rerurn Degree_info

3.1.2 Sentence Similarity Computation

The sentence similarity is calculated from pre-processed materials. Each phrase is analyzed and its degree of information, which is its weight based on term occurrence frequency, is calculated (eqn.3).

$$Deg_{inf_i} = OF(T_i) \times \log\left(\frac{n}{n_k}\right) \quad (3)$$

In the above equation, Deg_{inf_i} is the degree of information of i^{th} sentence concerning the term T_i. $OF(T_i)$ is the occurrence frequency of term t_k in i^{th} sentence. n_k is the total count of sentences that possesses t_k. The Inverse Sentence Frequency (ISF) in Vector Space Model (VSM) is represented by $\log\left(\frac{n}{n_k}\right)$. The inter-sentence similarity is then computed by the following equation.

$$SIM(S_i, S_j) = \frac{\sum_{k=1}^{m} Deg_{inf_{ik}} Deg_{inf_{jk}}}{\sqrt{\sum_{k=1}^{m} De\ ^2{}_{inf_{ik}} \cdot \sum_{k=1}^{m} Deg^2{}_{inf_{jk}}}} \quad (4)$$

The similarity between sentences $S_i\ and\ S_j$ with respect to Deg_{inf_i} is presented. The comparison is performed between a target sentence with all the remaining sentences and this process is repeated for all the sentences.

$$ROUGE_{scr}(s_i, s_j) = \frac{\sum_{i=1}^{n} count(n-gram, target)}{count(n-gram(j))} \quad (5)$$

The value of $'ROUGE_{scr}'$ lies between 0 and 1, where 0 indicates that the bigram is not present in the reference sentence and 1 otherwise. It is ensured that the $ROUGE_{scr}$ ranges between 0 and 0.6, such that the summary is made.

4. PERFORMANCE ANALYSIS

The publicly available Arxiv database is employed for evaluating the performance of the proposed work. The Arxiv dataset is built by the articles downloaded from pre-prints of arxiv repository and were converted to plain text with the help of Pandoc . The performance of the work is evaluated in aspects of content coverage, cohesion, and readability of the formed summary. The content coverage of the summary is represented by

$$C_{CVG} = Sim(s_i, O_i) \quad (6)$$

Here, O is the collection of sentences and O_i is the weighted average of sentences in a document. The similarity between s_i and O_i is computed to measure the significance of the sentence. Greater cohesion indicates the interconnection between the sentences is greater and is represented by

$$C_{COH} = 1 - Sim(S_i, S_j) \quad (7)$$

Readability is measured based on inter-sentence similarity, as computed in eqn. (4). The greater value of readability emphasises that the readability of the summary is high.

$$C_{RD} = Sim(S_i, S_j) \quad (8)$$

This work employs ROUGE-N for comparison.

Table 1. Performance comparison w.r.t ROUGE-N (F-measure) for Arxiv Database

Performance Metric	Feature Extraction Technique	Worst	Mean	Best
ROUGE-1	Degree of Information technique	32.7 %	38.2 %	41.61 %
	FBES technique	38.4 %	42.2 %	**44.71** %
ROUGE-2	Degree of Information technique	11.4 %	13.86 %	15.89 %
	FBES technique	12.83 %	16.84 %	**18.56** %

Table 2. Precision, Recall and F-measure Scores of proposed and existing methods

Performance Metric	Feature Extraction Technique	Precision	Recall	F-measure
ROUGE-1	Degree of Information technique	44.6 %	39 %	41.61 %
	FBES technique	47.3 %	42.4 %	**44.71** %
ROUGE-2	Degree of Information technique	17.3 %	14.7 %	15.89 %
	FBES technique	21.4 %	16.4 %	**18.56** %

In addition, the proposed work is compared with other performance measures such as sensitivity ($\mathcal{S}$), Positive Predictive Value (PPV) and Summary Accuracy (SA).
The attained results based on these performance measures are tabulated in Table 3.

Table 3. Performance evaluation w.r.t $\mathcal{S}$, PPV, SA

Performance Metrics/Feature extraction techniques	Degree of Information technique	FBES technique
$\mathcal{S}$	52 %	66 %
PPV	44 %	57.19 %
SA	97.51 %	99.14 %

The ROUGE-N score and other metrics show that the combination of degree of information and ROUGE-based feature extraction approaches performs better. Cohesion denotes sentence connectivity, which affects sentence and paragraph flow. Summary cohesion is usually checked using cosine similarity and average sentence similarity. The following equations calculate Flesch Kincaid Grade Level (FKGL), Coleman Liau (CL), and Automated Readability Index (ARI) to assess summary readability.

$$FKGL = 0.39 \times \left(\frac{Words}{Sentences}\right) + 11.8 \times \left(\frac{Syllables}{Words}\right) - 15.59 \quad (13)$$

$$CL = 5.89 \times \left(\frac{Characters}{Words}\right) - 0.3 \times \left(\frac{Sentences}{Words}\right) - 15.8 \quad (14)$$

$$ARI = 4.71 \times \left(\frac{Character}{Words}\right) + 0.5 \times \left(\frac{Words}{Sentences}\right) - 21.43 \quad (15)$$

The results obtained by the proposed work by varying feature extraction techniques are are shown in Table 4.

Table 4. Readability analysis

Performance Metrics/Feature extraction techniques	Degree of Information technique	FBES technique
Cohesive Score	4.8 %	5.2 %
FKGL	14.02 %	15.23 %
CL	25.25 %	25.89 %
ARI	19.12 %	21.17 %

5. CONCLUSIONS

Degree of information and ROUGE score are used to extract text summaries in this article. The degree of sentence information and ROUGE score determine sentence similarity. The average F-measure scores for ROUGE-1 and ROUGE-2 are 44.71 and 18.56. The results show that the proposed summarization technique works and can be improved to produce abstractive summaries.

REFERENCES

[1] L. Dong et al., "Unified language model pre-training for natural language understanding and generation," in Adv. Neural Inform. Process. Syst., 2019, pp. 13 042–13 054.

[2]K. M. Hermann et al., "Teaching machines to read and comprehend," in Adv. Neural Inform. Process. Syst., 2015, pp. 1693–1701.

[3] H. Van Lierde and T.W. Chow, "Query-oriented text summarization based on hypergraph transversals," Title: Inform. Process.Manag., vol. 56, no. 4, pp. 1317–1338, 2019.

[4] M. Mohamed and M. Oussalah, "SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis," Inform. Process.&Manag., vol. 56, no. 4, pp. 1356–1372, 2019.

[5] Radford Alec,Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, and Sutskever Ilya, "Language models are unsupervised multitask learners enhanced reader," OpenAI Blog, vol. 1, no. 8, 2019.

[6] Ma, T., Pan, Q., Rong, H., Qian, Y., Tian, Y., & Al-Nabhan, N. (2021). T-bertsum: Topic-aware text summarization based on bert. IEEE Transactions on Computational Social Systems, 9(3), 879-890.

[7] Elsaid, A., Mohammed, A., Fattouh, L., & Sakre, M. (2022). A Comprehensive Review of Arabic Text summarization. IEEE Access.

[8] Mahalakshmi, P., & Fatima, N. S. (2022). Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques. IEEE Access, 10, 18289-18297.

[9] Mao, Q., Li, J., Peng, H., He, S., Wang, L., Philip, S. Y., & Wang, Z. (2022). Fact-Driven Abstractive Summarization by Utilizing Multi-Granular Multi-Relational Knowledge. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1665-1678.

[10] Ma, B., Sun, H., Wang, J., Qi, Q., & Liao, J. (2021). Extractive Dialogue Summarization Without Annotation Based on Distantly Supervised Machine Reading Comprehension in Customer Service. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 87-97.

[11] Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M., & Kabir, M. M. (2021). A survey of automatic text summarization: Progress, process and challenges. IEEE Access, 9, 156043-156070.

[12] Jiang, J., Zhang, H., Dai, C., Zhao, Q., Feng, H., Ji, Z., & Ganchev, I. (2021). Enhancements of attention-based bidirectional lstm for hybrid automatic text summarization. IEEE Access, 9, 123660-123671.