Generic document summarization approach based on controlled stochastic sentence selection

Supriya Gupta[1], Aakanksha Sharaff[2], Naresh Kumar Nagwani[3]

[1,2,3] Department of Computer Science and Engineering,
National Institute of Technology Raipur, India.
Corresponding Author: Supriya Gupta
ORCID ID- 0000-0001-5151-136X
sgupta.phd2018.cs@nitrr.ac.in, asharaff.cs@nitrr.ac.in, nknagwani.cs@nitrr.ac.in

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May2021

Abstract:
In the new norm and cloud world era, online document generation has exponentially increased. The readers from different genres are unable to filter redundant information at a fast-paced rate. The research work is beneficial in raising awareness of utilizing online text summarization for distance learning among teachers, researchers, and students. It enables academia to quickly access concise and precise information from varied online sources. An efficient document summarization model reduces the read-time and improves information diversity; the paper presents an extractive summarization technique with a controlled stochastic sentence selection mechanism. The controlled stochastic limit is fine-tuned using TF, cosine, and Jaccard similarity measures. This unique sentence selection strategy is combined with a meta-heuristic approach to generate multiple solutions iteratively. The fitness of summary solutions is evaluated concerning the original document set producing the final summary. The various algorithms used for summarization are compared with the recommended model. The ROUGE-1 and ROUGE-2 values are empirically evaluated over DUC 2001, DUC 2002 datasets, which showcase an increase of 34.49% in Recall over the existing methods.

Keywords: Extractive text summarization, controlled stochastic selection, similarity measure.

1. Introduction

Recently, the users, trainers, and researchers face issues in accessing and extracting relevant information quickly from the massive set of available text data in digital format. The increasing e-learning modules, e-government digital archives, biomedical literature, legal documents, and news articles require automatic text summarization to produce shortened and crisp summary information. The online space is flooded with redundant details which need to be synthesized in logical excerpts which are easy to read and comprehend. Extractive text summarization can be used to pick the most salient information from the plethora of documents and help explore more specific data of the user's interest. The trainers and scholars can be empowered by automated text summarization tools to grasp and recollect significant concepts in the generic field of mathematics, computer science, biomedical science, social science, law, and journalism-related subjects.

Text pre-processing and sentence selection are one of the key aspects for reducing redundancy without compromising the essence of meaningful information in extractive summarization. The dissimilarity and similarity between sentences can be exploited for expressing information in a condensed form. The researchers have studied different meta-heuristic algorithms and various similarity techniques in the past, affecting the performance of text summarization. The existing methods do not focus on tuning threshold parameters, responsible for enhancing the uniqueness of sentences in summary. The proposed methodology optimizes the selection process of sentences using different similarity measures and preserves versatile information in the generated summary text.

The different sections in this work are structured as follows. Segment 2 gives a concise insight regarding the connected work done in the past for text synopsis. Segment 3 describes the recommended method, similarity measures, and scheme of implementation approach. Segment 4 explains the various experiments and parameter tuning and the assessment of evaluation metrics of different algorithms. Finally, segment 5 consists of conclusive observations and suggestions to additionally improve the similarity-based summarization framework.

2. Literature review

The extractive summarization techniques keep the document information as it is and retains only the most essential points. The authors have defined a hierarchical method for summarization that proves to be good for smaller document sizes but has a limited accuracy [1], which makes the technique unusable for real-time document summarizations. Saini N et al. estimated the similitude or disparity between sentences utilizing cosine similarity, WMD, and NGD [2]. Results outline the prevalence of the expressed MOBDE approach with arbitrary determination. A corpus-based method [3] is described, registering the closeness between concise messages of sentence length utilizing a calculation that assesses semantic data and word request data suggested in the sentences. Deep learning is another area of research that has come up in recent times. Authors claim to have developed a Restricted Boltzmann Machine (RBM) based algorithm [4] to improve accuracy of extractive summarization. The general precision of the RBM algorithm relies upon the report substance and how the network is trained based on the input features.

Semantic similarity and Automatic Document Summarization-based extraction are performed utilizing the neuro-fuzzy genetic algorithm [5]. Word net-analyzer is used for semantic computing similarity. Clustering is also known as one technique that can be used for summarization; the proposed novel system [6] straightforwardly creates bunches coordinated with sentence positioning outcomes. In this paper, three distinctive positioning capacities in a bi-type archive diagram are developed from the given report set. The ranking is applied separately based on initial k clusters. A comparative study of the Indian language-dependent semantic graph-based abstractive text summarization technique is provided [7]. For Bengali text [8], the authors have evaluated abstractive text summary using deep learning model and then measured the sentence similarity between human summary and machine have given summary using cosine/Jaccard similarity, word mover distance techniques. Another work uses the Marathi language [9] using a graph-based model, combined with the ROUGE features. A language-independent system [10] is mentioned, which uses ROUGE features, and also uses sentence re-ranking, and ROUGE can be used for both language-dependent and independent cases. Aspects and Comments are two different sides of extractive document summarization [11]; authors have conducted experimental studies that portray users are inclined towards customized rundowns that precisely mirror their interests.

Comments-oriented document summarization uses a novel multi-aspect-co-rank model [12] to achieve better performance. Tamiya S. et al. have used sentence embeddings with feed-forward neural network and centroid embeddings vector [13], which considers the setting of words in a sentence and permits catching sentence semantics and connections between sentences. Deepa Anand et al. [14] have introduced an information-driven semi-supervised methodology for legal text extractive summarization, which doesn’t require domain knowledge or feature generation utilizing various neural network designs. Marzieh Oghbaie et al. have recommended a unique pairwise comparison likeness measure between two documents [15], but it can be conveniently adapted to any vector type. The researchers wanted to study and investigate the similarity/dissimilarity measures and their effect on performance when used on extensive data collections for text summarization. In this work, the similarity measures are used with controlled stochastic selection for summarization, presented in the next segment.

3. Methods

The preliminary process of the planned strategy takes a single document as input. The pre-processing techniques are applied to the raw text for meaningful harvesting data. A unique controlled stochastic selection is introduced, which plays a vital role in an effective summary generation. The different similarity measures - cosine, Jaccard, and TF are used for sentence evaluation based on the defined fitness function. The output summary is generated and selected based on max fitness when all criteria are met. The effectiveness of a system-based summary from the proposed model is compared with the corresponding golden summaries using ROUGE evaluation.

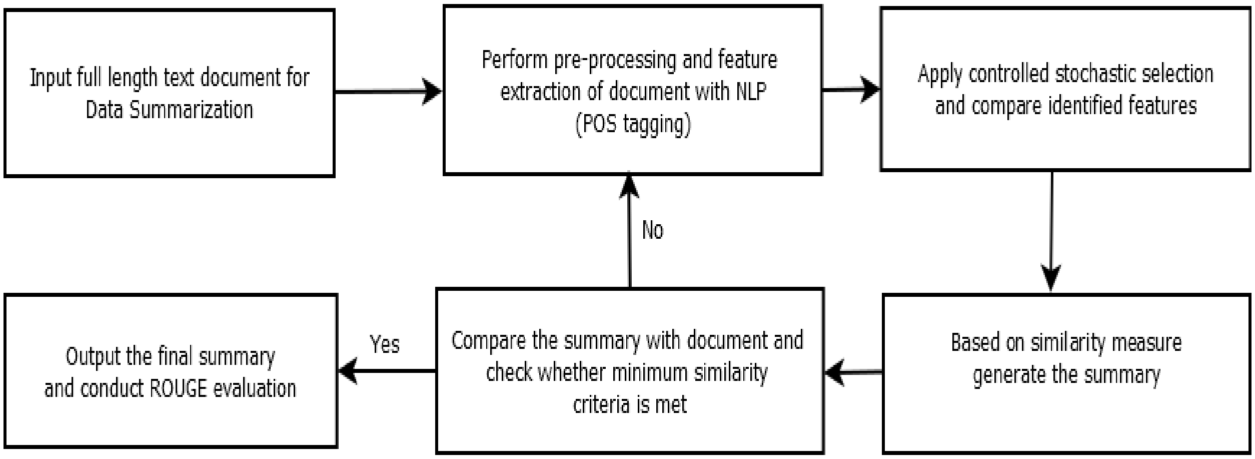

Input full length text document for Data Summarization
Perform pre-processing and feature extraction of document with NLP (POS tagging)
Apply controlled stochastic selection and compare identified features
No
Output the final summary and conduct ROUGE evaluation
Yes
Compare the summary with document and check whether minimum similarity criteria is met
Based on similarity measure generate the summary

Figure 1. The controlled stochastic extractive summarization process

3.1 Pre-processing

The process of extractive summary generation is depicted in Figure 1, where the input text document is picked in full-text form. This is further given to the feature extraction unit that applies critical language processing techniques, for example, evaluating n-grams, eliminating stop words, chunking, parts-of-speech labeling, and lemmatization in the text document.

3.2 Controlled stochastic metaheuristic approach

The proposed controlled stochastic selection-based metaheuristic approach is used to find the best solution using a mathematical model. The stochastic limit controls the selection of choosing dissimilar sentences in the individual summary candidates. The metaheuristic approach is used to generate multiple summaries initialized by the number of iteration and solutions. The generation of multiple summary candidates helps in obtaining the optimal fitness based on the non-redundancy feature.

3.3 Similarity/dissimilarity measure

A good summary length falls in the range of 30-50% of the actual text document. The summary length is configurable and, for the experiment, fixed at a 30% compression rate with the variation of the controlled stochastic limit. The choice of sentence candidates in the probable summary depended on the dissimilarity between them. The following similarity techniques are used for the evaluation of dissimilarity.

Cosine likeness is an extent of comparability between two vectors (non-zero) that calculates the cosine angle amid them. It will, in general, be portrayed like: The worth of this comparability lies between -1 to 1. Equation 1 shows two vectors are covering or precisely like one another, - 1 demonstrates two vectors are inverse to one another, and 0 shows they are symmetrical to one another [16]. As our records contain text, to quantify cosine closeness between two sentences, sentence vectors are required [17].

$$S_{cos}(s^1, s^2) = \frac{s^1\,(s^2)^T}{\sqrt{s^1\,(s^1)^T}\sqrt{s^2\,(s^2)^T}} \tag{1}$$

Jaccard similarity measures the closeness between two sets and is figured as the count of fundamental terms by the count of intriguing terms regarding the two sets (Jaccard, 1908), as shown in Equation 2. For our situation, set s^1 comprises the exceptional words in the main sentence, and set s^2 comprises the novel expressions of the subsequent sentence [18].

$$S_{jaccard}(s^1, s^2) = \frac{|\,s^1 \cap s^2|}{|\,s^1 \cup s^2\,|} \tag{2}$$

TF (Term Frequency) technique, as defined in Equation 3, makes sure that the overall document is described with the help of assistive features, which can be used in the other part of the process to help in better document summarization [19]. For example, if the input document has a lot of stop words like 'if,' 'and,' 'the,' etc. then, while comparing these words with the original document in the post-summary phase will make a lot of miscalculations due to the fact that the following two lines,

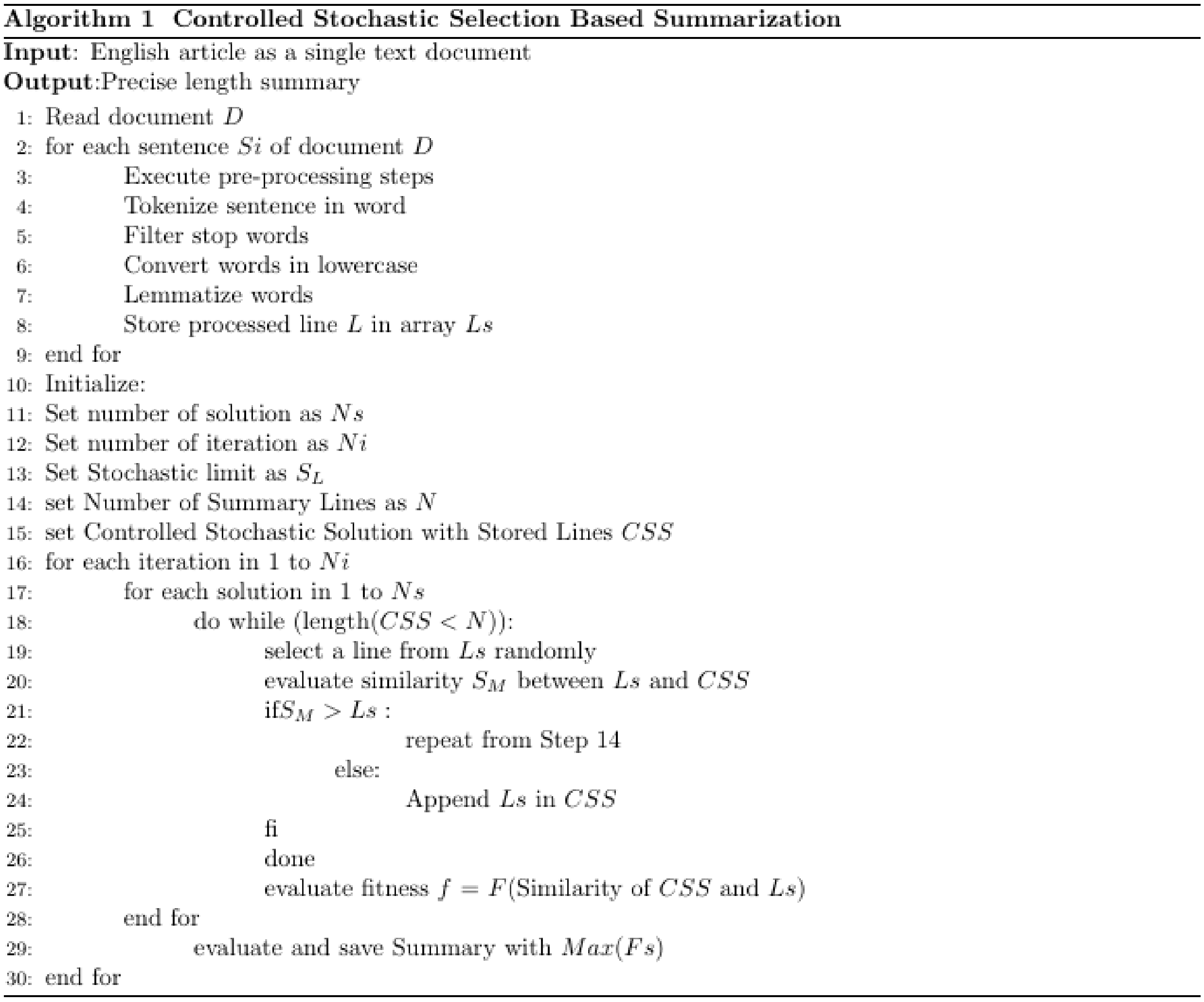
"We are going for a field trip."
and,
"We are having this for a wedding."

$$\mathrm{TF}(t, d) = \frac{f_d(t)}{\max f_d(w)} \tag{3}$$

Are entirely different sentences and have completely different meaning, but while comparison with each other, these lines will get a term frequency (TF) score of 0.57 for each line, which is incorrect. But, after application of NLP, POS tagging, and chunking for stop words removal, these lines will get transformed into,

"going field trip."
and
"having a wedding."

This will make the TF score of 0 for each of the lines, and thereby improving the accuracy of comparison, which in turn improves the accuracy of the overall summarization process at large. Thus, proper feature extraction from the input document is a must, and this block should be designed very carefully to achieve high accuracy for the summarizer.



Algorithm 1 Controlled Stochastic Selection Based Summarization

Input: English article as a single text document
Output:Precise length summary

```
1: Read document $D$
2: for each sentence $Si$ of document $D$
3:         Execute pre-processing steps
4:         Tokenize sentence in word
5:         Filter stop words
6:         Convert words in lowercase
7:         Lemmatize words
8:         Store processed line $L$ in array $Ls$
9: end for
10: Initialize:
11: Set number of solution as $Ns$
12: Set number of iteration as $Ni$
13: Set Stochastic limit as $S_L$
14: set Number of Summary Lines as $N$
15: set Controlled Stochastic Solution with Stored Lines $CSS$
16: for each iteration in 1 to $Ni$
17:         for each solution in 1 to $Ns$
18:                 do while (length($CSS < N$)):
19:                         select a line from $Ls$ randomly
20:                         evaluate similarity $S_M$ between $Ls$ and $CSS$
21:                         if$S_M > Ls$ :
22:                                         repeat from Step 14
23:                                 else:
24:                                         Append $Ls$ in $CSS$
25:                         fi
26:                         done
27:                         evaluate fitness $f = F$(Similarity of $CSS$ and $Ls$)
28:         end for
29:                 evaluate and save Summary with $Max(Fs)$
30: end for
```

The above Algorithm 1 depicts that the stochastic process is controlled using the stochastic similarity limit parameter s_L; this parameter is selected by the user on run-time. The proposed algorithm is explained below with relevant equations:

Let the full-text document is represented by D where;

$$D = (L1, L2 \dots \dots . Ln) \tag{4}$$

D consists of n number of tokenized lines starting from $L1, L2, \dots \dots \dots to\ Ln$ as formulated in the above Equation 4.

$$CSS\ = \sum_{k=1}^{N} (\sum_{i=1}^{n} \sum_{j\neq i}^{n} (1 - sim)(Li, Lj) > s_L) \tag{5}$$

The above Equation 5 defines the controlled stochastic solutions CSS based on the non-redundancy feature, where n represents the total count of lines present in the D document, N is the compression factor, and s_L represents the stochastic limit defined to control the degree of dissimilarity between sentences compared within the document.

3.4 Optimal summary selection

Multiple summary solutions are generated by the above metaheuristic algorithm, which is further scrutinized based on the fitness function explained below.

$$f = \sum_{k=1}^{N} (sim\ (CSS, D - CSS)) \tag{6}$$

The fitness function f is defined in the above Equation 6, which provides the relation between the solution summaries and the original document where N represents the count of lines present in solution candidates with the compression rate. This Fitness function is calculated by finding similarities between the controlled stochastic solution and the rest of the document.

$$Final\ summary = \max(f)\ \ summary\ solutions \tag{7}$$

The other solution summaries are generated by the summarization model, and as per the above Equation 7, the solution with maximum fitness gets qualified for the ultimate summary.

4. Results and observations

4.1 Experimental Setup

The benchmark datasets, DUC-2001 having 309, and DUC-2002 having 567 text news reports (as archives), individually written in English, are utilized for the text summarization experiments as depicted below in Table 1. A golden summary of approximately 100 words is available for each record [20]. The golden summary is used distinctly for the assessment of the created summary.

Table 1. DUC 2001 and DUC 2002 Dataset description

Dataset Description	**DUC 2001**	**DUC 2002**
Origin	TREC	TREC
Count of Documents	309	567
Themes	30	59
Summary Length	100 terms	100 terms

4.2 Evaluation matrices

ROUGE analysis is used to evaluate the recommended summarization framework [21]. Through this analysis, overlapping terms are assessed over the golden and proposed model summaries. Elevated ROUGE value gives more closeness of generated summary with the golden summary. ROUGE score is defined below in Equation 8:

$$ROUGE - N = \frac{\sum s \in summary_{actual} \sum N{-}grams \in S\ Count_{match}\ (N{-}gram)}{\sum S \in summary_{actual} \sum N{-}grams \in S Count\ (N{-}gram)} \qquad [8]$$

N addresses the n-gram length, $Count_{match}(\ N - gram)$ represents the greatest count of covering $N - grams$ among golden and evaluated solution, $Count\ (N - gram)$s means the all outnumber of $N - grams$ found in the golden solution. ROUGE1 considers the count of unigrams, and ROUGE2 represents the count of bi-grams in the golden and evaluated summary [29].

4.3 Parameter setting

The study of the stochastic selection of sentences is carried out with cosine, Jaccard, and TF similarity measures calculations. The various input parameters like number of iterations, number of solutions, compression rate, controlled stochastic limit, and usage of specific similarity measure are fine-tuned for finding the salient sentences to form summary solutions. The selection of dissimilar sentences with stochastic limits SL has improved the non-redundant information in summary. In the proposed method, multiple values 0.60, 0.70, and .80 are used to understand the impact of stochastic limit S_L selection.

4.3.1 Compression rate

The proposed model provides flexibility in choosing the optimal compression rate for text summarization. It is observed that efficient results are obtained when the compression rate is set at 30% of the full-text document. It can be used for the generation of precise and custom length summaries according to the user's input criteria.

4.3.2 Controlled stochastic limit

In the proposed method, the metaheuristic approach is used to find the optimal solutions based on maximizing the fitness function. The distribution of various ROUGE scores with individual similarity measures are displayed in Table 2. After multiple trials, a stochastic limit of 0.7 is found appropriate to improve the diversity in the final summary. The best ROUGE values are observed when S_L is set at 70%, which governs the uniqueness of sentences in the summary solution. The compression ratio for the proposed algorithm is controlled by the output parameter N, which is varied according to the user. N is kept at 30% of the input lines, which makes the algorithm's output to be adaptive in terms of the actual number of output lines.

Table 2. Stochastic controlled limit at 30% compression with ROUGE score of DUC dataset

Dataset	Stochastic limit	Compression rate	Similarity measure	ROUGE-1	ROUGE-2	ROUGE-SU4	ROUGE-L
DUC 2001 Document 5089 Summary for 100 words	.8	30%	Cosine	0.62385	0.18447	0.29897	0.44
			Jaccard	0.59633	0.19417	0.28454	0.42
			TF	0.58322	0.16505	0.2701	0.36
	.7	30%	Cosine	0.65138	0.27184	0.34021	0.46667
			Jaccard	0.55046	0.17476	0.25773	0.32
			TF	0.55963	0.14563	0.25155	0.38667
	.6	30%	Cosine	0.58716	0.20388	0.28454	0.4
			Jaccard	0.59633	0.19417	0.27629	0.42667
			TF	0.55963	0.12621	0.24124	0.33333
DUC 2002 Document 0060 Summary for 100 words	.8	30%	Cosine	.055963	0.19608	0.28632	0.4444
			Jaccard	0.53211	0.11765	0.21474	0.4244
			TF	0.55046	0.16667	0.26737	0.4103
	.7	30%	Cosine	0.63303	0.29529	0.38263	0.56944
			Jaccard	0.59633	0.16667	0.26105	0.45833
			TF	0.5674	0.27451	0.35789	0.54167
	.6	30%	Cosine	0.6055	0.28431	0.36632	0.51389
			Jaccard	0.55963	0.23529	0.31158	0.47222
			TF	0.56881	0.26471	0.32842	0.48611

The performance of Average Recall values across DUC-2001 and DUC-2002 document collection and the ROUGE evaluations over various stochastic limits with multiple similarity techniques is represented in Figure 2.

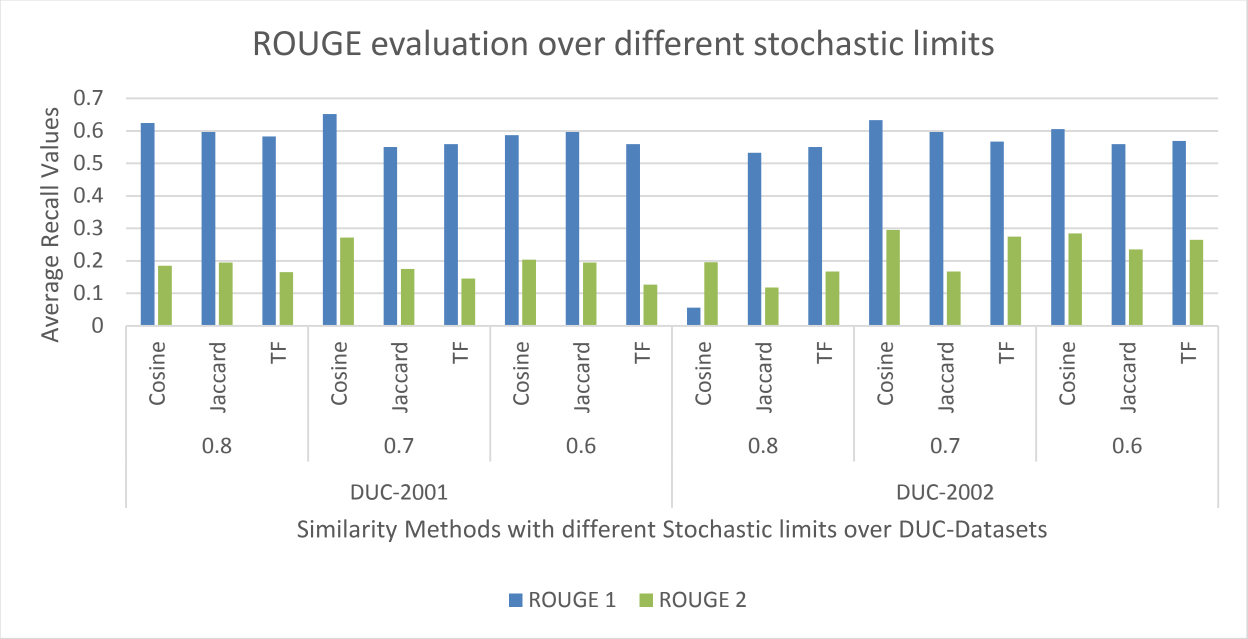


Figure 2. ROUGE evaluation over different stochastic limits

4.4 Comparison of methods

The performance of the proposed algorithm with DUC datasets and some of the manually generated datasets are compared. It is observed that the changes in the input dataset become irrelevant when a large number of documents are taken for summarization. The cosine similarity measure outperformed the other two Jaccard and TF similarity techniques used for summarization. The various parameter values are initialized to specific values to get the maximum fitness scores while performing the experiments. The different methods which are compared for the extractive text summarization are listed in Table 3.

Table 3. ROUGE of average Recall comparison between various algorithms

Algorithms	**DUC-2001**		**DUC-2002**	
	ROUGE-1	**ROUGE-2**	**ROUGE-1**	**ROUGE-2**
DE [22]	.4786	.1853	.4669	.1237
Cosum [23]	.4727	.2012	.4908	.2309
Unified Rank [24]	.4538	.1765	.4849	.2146
SVM [25]	.4463	.1702	.4324	.1087
FEOM [26]	.4773	.1855	.4658	.1249
Manifold ranking [27]	.4336	.1664	.4233	.1068
QCS [28]	.4485	.1852	.4487	.1877
Proposed Controlled stochastic selection based summarization	**.5940**	**.2685**	**.6601**	**.2850**

The average Recall values obtained from these methods QCS (Query cluster summarize), FEOM (Fuzzy evolutionary optimization model), SVM (support vector machine), DE (Differential evolution), are compared with the proposed technique in Figure 3. The results and analysis indicate that the recommended model is at par compared with the other existing methods.

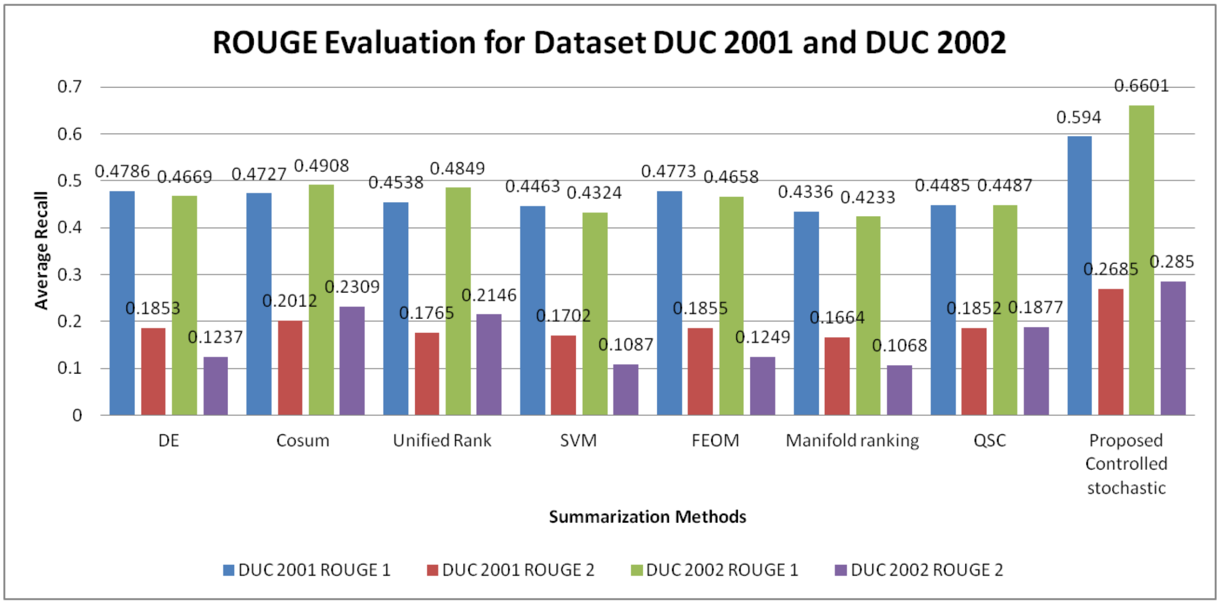


Figure 3. ROUGE evaluation of proposed method vs. existing methods

5. Conclusion

In the digital setup offered to instructors and researchers working on multiple domains, without summarization, repetitiveness can deteriorate overall summary quality. In the proposed work, the non-redundancy of information acts as a crucial feature in selecting salient sentences in the summary document. The dissimilarities between sentences are evaluated using different similarity measures in a controlled manner by using a controlled stochastic limit. The proposed algorithm can accurately find out the repetitive sentences, and remove them from the input document, thereby giving a very accurate extractive summary. This method performed well on the moderately sized documents taken from the standard datasets (DUC-2001 and DUC-2002). The presented work is found at par in comparison to the other existing methods using ROUGE evaluation. The unique controlled-stochastic sentence selection-based text summarization method would keep the computing and mathematics enthusiasts interested in effectively carrying out their research and innovation work. In the future, this model can be extended for multi-document summarization with different types of similarity measures to improve performance.

References

[1] H. Xu, Z. Wang and X. Weng. (2019).Scientific Literature Summarization Using Document Structure and Hierarchical Attention Model, IEEE Access, vol. 7, pp. 185290-185300.

[2] Saini N, Saha S, & Chakraborty D Bhattacharyya. (2019). Extractive single document summarization using binary differential evolution: Optimization of different sentence quality measures. PLOS ONE 14(11): e0223477.

[3] Li Yuhua, McLean David Bandar, Zuhair, O'Shea, James & Crockett, Keeley. (2006). Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering. 18 1138-1150. 10.1109/TKDE.2006.130.

[4] PadmaPriya, G. & K. Duraiswamy. (2014).An Approach for Text Summarization using Deep Learning Algorithm. JCS 10: 1-9.

[5] K. Srinivasa Rao, D.S. R. Murthy,& Gangadhara Rao Kancherla. (August 2019) .Semantic Similarity Based Automatic Document Summarization Method. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6.

[6] X. Cai & W. Li. (July 2013). Ranking Through Clustering: An Integrated Approach to Multi-Document Summarization. IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 7, pp. 1424-1433.

[7] C. Sunitha, A. Jaya, & Amal Ganesh. (2016). A Study on Abstractive Summarization Techniques in Indian Languages. Procedia Computer Science, Volume 87, Pages 25-31, ISSN 1877-0509.

[8] Masum, Abu Kaisar Mohammad, Sheikh Abujar, Raja Tariqul Hasan Tusher, Fahad Faisal & Syed Akther Hossain. (2019). Sentence Similarity Measurement for Bengali Abstractive Text Summarization. 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 1-5.

[9] Dakulge, Umakant, Dharmadhikari, & S. C.(May 2014). Automated Text Summarization: A Case Study for Marathi Language. Data Mining and Knowledge Engineering, [S.l.], v. 6, n. 3, p. 99-105, ISSN 0974 – 9578.

[10] Dhawale A.D., Kulkarni S.B.,& Kumbhakarna V. (2020). Survey of Progressive Era of Text Summarization for Indian and Foreign Languages Using Natural Language Processing. Innovative Data Communication Technologies and Application. ICIDCA 2019.vol 46. Springer, Cham.

[11] Berkovsky S., Baldwin T., & Zukerman I. (2008). Aspect-Based Personalized Text Summarization. In: Nejdl W., Kay J., Pu P., Herder E. (eds) Adaptive Hypermedia and Adaptive Web-Based Systems. AH, 2008. Lecture Notes Computer Science, vol 5149. Springer, Berlin, Heidelberg.

[12] Huang L., Li H., Huang L. (2013). Comments-Oriented Document Summarization Based on Multi-aspect Co-feedback Ranking. In: Wang J., Xiong H., Ishikawa Y., Xu J., Zhou J. (eds) Web-Age Information Management. WAIM 2013. Lecture Notes in Computer Science, vol 7923. Springer, Berlin, Heidelberg.

[13] Lamsiyah S., El Mahdaouy A., El Alaoui S.O.,& Espinasse B. (2020). A Supervised Method for Extractive Single Document Summarization Based on Sentence Embeddings and Neural Networks. In: Ezziyyani M. (eds) Advanced Intelligent Systems for Sustainable Development (AI2SD'2019). AI2SD 2019. Advances in Intelligent Systems and Computing, vol 1105. Springer, Cham.

[14] Deepa Anand, & Rupali Wagh. (2019). Effective deep learning approaches for summarization of legal texts.Journal of King Saud University - Computer and Information Sciences, ISSN 1319-1578.

[15] Marzieh Oghbaie, & Morteza Mohammadi Zanjireh.(2018). Pairwise document similarity measure based on present term set. Journal of big data.

[16]Aliguliyev RM. A new sentence similarity measure and sentence-based extractive technique for automatic text summarization. Expert Systems with Applications. 2009; 36(4):7764–7772.http://doi.org/10.1016/j.eswa.2008.11.02

[17] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.

[18] Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. M. B. (2004). The similarity metric. IEEE Transactions on Information Theory, 50(12), 3250–3264. https://doi.org/10.1109/TIT.2004.83810

[19] Eminagaoglu, M. (2020). A new similarity measure for vector space models in text classification and information retrieval. Journal of Information Science. https://doi.org/10.1177/0165551520968055

[20] (https://www-nlpir.nist.gov/projects/duc/data.html0

[21] Lin CY. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out. 2004.

[22]Aliguliyev RM. A new sentence similarity measure and sentence-based extractive technique for automatic text summarization. Expert Systems with Applications. 2009; 36(4):7764–7772. https://doi.org/.1016/j.eswa.2008.11.022

[23] Alguliyev RM, Aliguliyev RM, Isazade NR, Abdi A, Idris N. COSUM: Text summarization based on clustering and optimization. Expert Systems. 2018; p. e12340.

[24] Wan X. Towards a unified approach to simultaneous single-document and multi-document summarizations. In: Proceedings of the 23rd international conference on computational linguistics. Association for Computational Linguistics; 2010. p. 1137–1145.

[25] Yeh JY, Ke HR, Yang WP, Meng IH. Text summarization using a trainable summarizer and latent semantic analysis. Information processing & management. 2005; 41(1):75–95. https://doi.org/10.1016/ j.ipm.2004.04.003

[26] Song W, Choi LC, Park SC, Ding XF. Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization. Expert Systems with Applications. 2011; 38(8):9112–9121. https://doi.org/10.1016/j.eswa.2010.12.102

[27] Wan X, Yang J, Xiao J. Manifold-Ranking Based Topic-Focused Multi-Document Summarization. In: IJCAI. Vol. 7; 2007. p. 2903–2908.

[28] Dunlavy DM, Oaˆ Leary DP, Conroy JM, Schlesinger JD. QCS: A system for querying, clustering, and summarizing documents. Information processing & management. 2007; 43(6):1588–1605. https://doi. org/10.1016/j.ipm.2007.01.003

[29] Liana Ermakova, Jean-Valère Cossu, Josiane Mothe: A survey on evaluation of summarization methods

, April 2019 Information Processing & Management 56(5) DOI:10.1016/j.ipm.2019.04.001

[30] Oliveira H, Ferreira R, Lima R, Lins RD, Freitas F, Riss M and Simske S J 2016 Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. Expert Systems with Applications 65: 68–86

[31] Abbasi-ghalehtaki R, Khotanlou H and Esmaeilpour M 2016 Fuzzy evolutionary cellular learning automata model for text summarization. Swarm and Evolutionary Computation 30: 11–26

[32] Alguliev R M, Aliguliyev R M, Hajirahimova M S, and Mehdiyev C A 2011 MCMR: maximum coverage and minimum redundant text summarization model. Expert Systems with Applications 38: 14514–14522

[33] Rautray R and Balabantaray R C 2017 Cat swarm optimization based evolutionary framework for multi-document summarization. Physica A: Statistical Mechanics and its Applications 477: 174–186

[34] Verma P and Om H 2019 MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. Expert Systems with Applications. 120: 43–56

[35] Ray R L 2010 Introduction to information retrieval. Journal of the American Society for Information Science and Technology 4: 852–885