

Particle swarm optimization model for Hindi text summarization

Rekha Jain *

Linesh Raja [†]

Department of Computer Applications

Manipal University Jaipur

Jaipur 303007

Rajasthan

India

Sandeep Kumar Sharma [§]

Department of Computer and Communication Engineering

Manipal University Jaipur

Jaipur 303007

Rajasthan

India

Devershi Pallavi Bhatt [‡]

Department of Computer Applications

Manipal University Jaipur

Jaipur 303007

Rajasthan

India

Abstract

Text Summarization is one of the techniques that shorten the original text without vanishing its information as well as meaning. A lot of algorithms exist for text summarization. Two approaches namely Abstractive Text Summarization and Extractive Text Summarization are used for this purpose. In Abstractive text summarization, the entire document is regenerated using a few lines. Whereas in Extractive Text Summarization sentences are filtered based on some ranks assigned to them by a specific algorithm. A lot of work has already been done in languages like English, Chinese etc. In this paper, the authors

* E-mail: rekha.jain@jaipur.manipal.edu (Corresponding Author)

[†] E-mail: linesh.raja@jaipur.manipal.edu

[§] E-mail: sandeep.sharma@jaipur.manipal.edu

[‡] E-mail: devershipallavi.bhatt@jaipur.manipal.edu

propose the summarization of Hindi text using the Particle Swarm Optimization model. Initially, the text in the Hindi language is summarized using a ranking-based technique then PSO (Particle Swarm Optimization) is applied to have an optimized summary of the text. One of the ranking-based techniques i.e. TF-IDF is introduced. Implementation of the proposed Systems is initially discussed in five steps- preprocessing, feature extraction, ranks generation, post-processing and optimized summarization using PSO. At the end, results are shown in terms of an optimized summary of text in a specific language. This system can be implemented in any standard language, but Hindi is selected for practical implementation because very few research work is done in the Hindi Language.

Subject Classification: Primary 93A30, Secondary 49K15.

Keywords: PSO model, Text summarization, Extractive text summarization, Ranking based techniques, Optimized text summarization, Hindi, TF-IDF.

1. Introduction

Automatic summarization of text is the most challenging task, as when a human summarizes the text, he/she reads the whole contents of the text so a good level of understanding can be developed. Since computers lack human knowledge and intelligence so it makes this task very difficult and challenging [1]. In today's era, the Internet is overflooded with information and it parades a lot of challenges in the field of Information Retrieval. It raises the need for an automated system that can summarize the text without vanishing its meaning. The proposed system summarizes the text in an optimized manner using a ranking-based approach with the PSO technique. This process summarizes the text in a manner that preserves its meaning intact, eliminating redundant information from the text and providing concise information to the user. This system is based on Unsupervised Machine Learning, where no training takes place in advance. Text is directly fed to the system and the system performs various tasks and finally, a summary is produced. Section 2 describes types of text summarization with current research in Text Summarization, and Section 3 focuses on proposed methodology. Section 4 shows the results. This system can be used in various applications viz news article summarization, search engines results summarization etc. Initially, text is broken into sentences then various preprocessing steps are performed like parsing, stop word removal, and root word identification. After it, a model is created for feature extraction. A specific algorithm is applied to calculate the score of text elements and generate the various versions of summaries. At the end, among all the generated summaries an optimized summary is fetched using the PSO technique.

2. Literature Survey

Types of Text Summarization

Techniques of summarizing the text can be classified according to Figure 1[2].

Based on input type [2]:

- Single Document: In this method input to the system is only one document. The system summarizes the text of a single file.
- Multi-Documents: Here input to the system is comparatively long, consisting of multiple documents.

Based on Output type [3]:

- Extractive: In this method of summarization important sentences are filtered from input text that are candidates for summary. Most approaches are based on this method. This method reads the text, assigns importance to sentences according to relevance and understanding and finally generates the summary according to importantly ranked sentences. The proposed system is based on Extractive Text Summarization.
- Abstractive: This method generates the summary like a human being as sentences are rephrased in this type of summarization. This method tries to guess the meaning of the text with the help of important phrases and generates the new text by putting them together in a meaningful and different manner. It is more attractive but difficult to achieve as well as an expensive task [4].

Based on Purpose [3]:

- Generic: In this method no assumptions about the domain specific knowledge are considered. All the documents are treated as homogenous [3].
- Domain Specific: Summary created by this model is much accurate as it takes domain knowledge into consideration [3].
- Query Based: Here summaries have only answers of queries executed by the system upon text [3].

A lot of work has been done in the field of summarization, and most of the research is done in either an extractive or abstractive manner. The

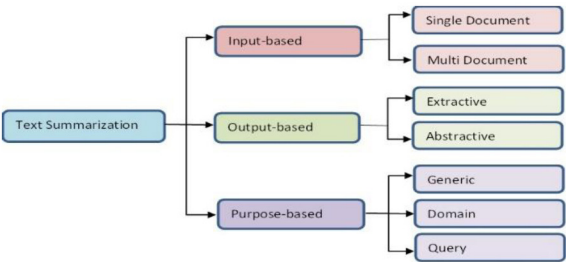


Figure 1
Types of Text Summarization

current paper proposes extractive text summarization. Table 1 concludes the research in recent years related to text summarization [5].

Table 1
Current research in Text summarization

Author, Year of publication	Technique/ Evaluation	Features	Performance
Filho, B. et al, 2007 [6]	Linguistic Approach / ROUGE	Focused on Sentence ordering and position	Evaluation parameters were precision, recall and f-measure through ROUGE metrics.
Makoto, H. et al, 2007 [7]	Linguistic Approach / (ROUGE-1, ROUGE-2 and ROUGE- W)	Describes Semantic relation between sentences	Semantic relation measures were used in bi-relation and parallel connection models.
Chen, Y.T et al, 2008 [8]	Machine learning / ROUGE metric.	Lexical, Prosodic, Confidence and Relevance features were extracted	Chinese broadcastnews was summarized using different models.
Xie, S.S. et al, 2009, [9]	Machine Learning (Semi-Supervised Learning) / ROUGE metric.	Focused on Local context information, its topic and speakers.	Prosodic and non-prosodic information was measured.
Araly, B. et al, 2011 [10]	Statistical Approach / ROUGE metric	Sentence scores were calculated by TextRank algorithm	Performance was measured for SynSem and on the DUC 2002

Mendoza, M. et al, 2014 [11]	Machine Learning / ROUGE-1, ROUGE-2	Engrossed on the title, length, position as well as Cohesion parameters	Performance was measured. through two Datasets DUC 2001 and 2002
Yadav, J. et al, 2016 [12]	Linguistic approach / ROUGE	Described, title similarity, sentence centrality,	Performance was evaluated for 95 documents through ROUGE-N. precision, recall and f-measure calculated.
Nallapati, R. et al, 2017 [13]	Machine Learning Approach / ROUGE (R-N, R-L)	Focused on Information Content, Salience, and Novelty.	Different dataset were measured using ROUGE.
Gupta, S et al, 2018 [14]	Hybrid Approach / ROUGE measuring precision, recall and f- score.	Introduced unique-terms, Title-Similarity, Numeric Token, Positive-Negative, keywords, Sentence length	Page Rank Algorithms, MMR algorithms were used and compared.
Begum, M. et al, 2020 [15]	Machine Learning (Supervised learning) (ROUGE-1,2)	Sentence Ranking is used in summary generation	LSTM-NN based summarization methods are used and evaluated
Darshna, P. et al, (2022) [16]	Machine Unsupervised	Sentences scores help to generate the summary	English as well as hindi text was summarized using graph based techniques
Liu, B., (2023) [17]	Machine Learning / Rouge, Bleu	Sentences ranked as well as reformed	Both extractive and abstractive summarization is done on same data set.

All the mentioned research generates the summarized text but is not guaranteed for optimal summarization. For this, the authors propose Particle Swarm Optimization (PSO) algorithm. This algorithm is implemented as a layer-on ranking-based technique and generates an optimized summary of the text.

3. Proposed Methodology

3.1 Preprocessing

Before passing the text to algorithm some kind processing needs to be done as algorithm can understand only mathematical form not the textual one. For this, following steps need to be done-

1. Converting the text into lower case
2. Dividing the text into sentences (sentence tokenization)
3. Dividing the sentences into words.
4. Removal of stop words
5. Stemming of words (converting them into root form)
6. Creating vector representation of words

3.2 *Ranking-Based Technique*

Initially Ranking Based techniques are used in citation analysis of web pages and analysis of the link structure of WWW. Various algorithms exist that follow this technique like the PageRank algorithm, and HITS algorithm. These algorithms help to compute and assign the rank to web pages. These pages are the part of search result according to the rank produced by the algorithm. Highly ranked pages appear at the top of search results. A similar kind of concept can be applied to Natural Language Processing documents resulting in summarized text. Various algorithms exist for this task e.g., TextRank algorithm [18], TF-IDF algorithm [19]. These algorithms are used to draw the knowledge from the entire text that helps them to mark ranking to sentences. These ranking-based algorithms can be represented in terms of undirected graphs. A graph is a collection of vertices and edges where vertices are denoted by sentences and edges are denoted by the connection between sentences [20]. This concept can be applied to Natural Language Processing where lexical and semantic graphs help us to draw knowledge from entire text with the help of a ranking mechanism [21][22]. The graph-Based Ranking algorithm helps us to decide the importance of the vertex in the graph based on the structure of the graph. In the proposed work TF-IDF algorithm is used for rank generation. Vector is passed to algorithm and algorithm calculates Term Frequency and Inverse Document Frequency of words. Then sentence's TF-IDF score is calculated based on the words' TF-IDF score. Ranks are assigned to the most important sentences. The selection of sentences is done based on a threshold value. Finally, a summary is generated. Mathematical formula for TF-IDF is discussed in next section.

3.3 *Mathematical Formulation*

Here a mathematical model is introduced for ranking algorithms i.e., TF-IDF ("Term Frequency-Inverse Document Frequency"). It is based on a

numeric measure that helps to provide ranks to sentences. This algorithm consists of two sub-modules which are multiplied together. These are as follows- [23][24][25].

- TF: It represents occurrences of a word in a document divided by total no of words in the document.

$$TF(t) = \frac{\text{No of times term } t \text{ appears in a doc}}{\text{Total no of terms in the doc}} \quad (1)$$

- IDF: It represents the uniqueness or rareness of a word.

$$IDF(t) = \log_e \frac{\text{Total number of documents}}{\text{No of documents with term } t} \quad (2)$$

hence TF-IDF can be calculated as-

$$TFIDF(t) = TF(t) * IDF(t) \quad (3)$$

3.4 Particle Swarm Optimization

The Particle Swarm Optimization (PSO) algorithm is used to solve optimization problems. This algorithm is based on the simulation of the social behavior of birds [26][27][28]. In swarm theory PSO is used in bird flocking, fish schooling etc. In PSO each solution is treated as a particle with a specific velocity to fly in the problem space. This algorithm assumes, “knowledge is shared among all the members of the same generation”. This technique optimizes the problem in an iterative manner. It starts with a population (solution candidate), which is called swarm particles. Each particle knows its global best (gbest) position within the swarm. Its own best position is calculated so far using the search space process. In each iteration particle’s velocity and position are updated that help to decide the optimal value. Iterations are repeated till some stopping criterion. The best outcome for each particle is known as pbest. gbest is the best value among all the particles. Next section describes the PSO algorithm.

Algorithm 1: PSO algorithm

X^i : position of particle or agent, V^i : velocity of particle or agent,
 $r1, r2$: random numbers between 0 and 1
 w : inertia weight, c_1 : cognitive constant
 c_2 : social constant, pbest: personal Best, gbest: global Best

Step 1. Generate initial population of particles (agents) evenly distributed over X .

Step 2. Evaluate the position of each particle as

$$X^i(t) = (x^i(t), y^i(t)) \quad (4)$$

$$pbest^i = X^i(t) \quad (5)$$

Step 3. Evaluate the velocity of each particle as

$$V^i(t) = (v_x^i(t), v_y^i(t)) \quad (6)$$

Step 4. Update the velocity of each particle as

$$V^i(t+1) = wV^i(t) + c_1 r_1 (pbest^i - X^i(t)) + c_2 r_2 (gbest - X^i(t)) \quad (7)$$

Step 5. Calculate pbest for each particle.

$$pbest^i(t+1) = pbest^i(t) + V^i(t) \quad (8)$$

Step 6. Update pbest i.e. If a particle's current position is better than its previous best position.

Step 7. Update gbest i.e., Find the best particle (according to the last best places of particles: best(pbest)).

$$gbest = \max(pbest^i(t)) \quad (9)$$

Step 8. check the stopping criterion if false Go to step 4.

This research focuses on extractive summarization of text using ranking based mechanism. This mechanism finds most relevant sentences from pool of sentences. It is an unsupervised technique that does not require labelled data set. Figure 2 shows the flow of work. The similarity

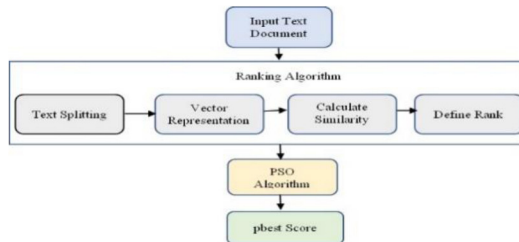


Figure 2

Proposed Methodology

index or sentence score of the text has been defined using the ranks, which is again calculated by PSO algorithm in order to generate the optimized summary. In this research work Hindi news data set is used for experimental purpose. Various versions of summaries are produced by TF-IDF algorithm according to various compression ratios. Finally, all the generated summaries are passed to PSO, and optimized summary is fetched based on gbest and pbest values.

3. Experimental Results

For the experimental purpose Hindi Dataset for Health is taken from Kaggle [29]. Firstly, this dataset is preprocessed means splitting of sentences is performed, stop word removal & root word identification is done. Then this preprocessed data is passed to the system which follows TF-IDF algorithm. This algorithm worked in iterative manner and produced various versions according to compression ratio. These were 20%, 40%, 50%, 60% and 70%. Now all the versions of summarized text are fed to PSO to fetch most optimized summary of text. PSO algorithm computes the pbest values for each summary in iterative manner with various parameters until the value converges. Highest pbest value is treated as gbest (global best) value for all the summaries. Finally, summary that contains highest pbest is considered as optimized summary.

Normally Bleu and Rouge scores are used to evaluate the generated candidate Summary in abstractive manner. In the experimental work a layer is implemented to find the optimized one among all the generated summaries in extractive manner. From above Figure 3, it can be concluded that the summary with 70% compression ratio is an optimized summary. Accuracy of this summary is calculated in terms of highest pbest value

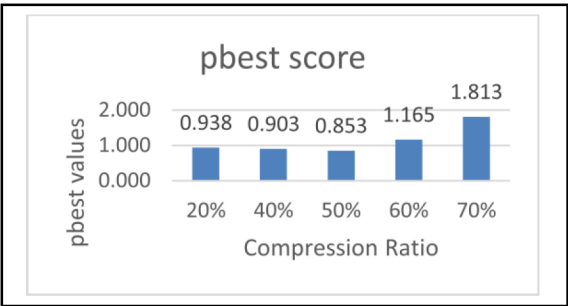


Figure 3
pbest scores for various versions of summaries

which is equal to gbest value. In given data set 70% compression ratio is considered as highly accurate summary among all the generated summaries as this summary contains highest pbest value that is equivalent to gbest.

5. Conclusion & Future Directions

Evaluation of produced summary is one of the challenging tasks as it is subjective matter. It is expected that produced summary should be concise, meaningful as well as it should not vanish the meaning. Also, it must be in optimal manner. In above mentioned algorithm layered technique of PSO is applied through TF-IDF algorithm. In future authors would compare this algorithm on various parameters like accuracy, precision, recall etc. Algorithm efficacy would be compared.

References

- [1] Dr. Michael J. Garbade, "A Quick Introduction to Text Summarization in Machine Learning", at towardsdatascience.com (2018).
- [2] Kushal Chauhan, Unsupervised Text Summarization using Sentence Embeddings, (2018).
- [3] Rekha Jain, Automatic Text Summarization of Hindi Text Using Extractive Approach, ECS Transactions, pp 4469-4477 (2022).
- [4] Khatri, Chandra and Singh, Gyanit and Parikh, Nish, Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks, (2018).
- [5] Avaneesh Kumar Yadav, Ashish Kumar Maurya, Ranvijay, Rama Shankar Yadav, "Extractive Text Summarization Using Recent Approaches: A Survey", (2021).
- [6] Filho, B., Paulo, P., Salgueiro Pardo, T.A., das Gracas Volpe Nunes, M. Summarizing scientific texts: Experiments with extractive summarizers. pp. 520-524 (2007).
- [7] Makoto, H., Shinnaka, Y., Iwano, K., Furui, S. Sentence-extractive automatic speech summarization and evaluation techniques. pp 1151-1161 (2006).
- [8] Chen, Y.T., Chen, B.L., Wang, H.M. A Probabilistic Generative Framework for Extractive Broadcast News Speech Summarization. pp 95-106 (2008).

- [9] Xie, S.S., Hakkani-Tür, D., Favre, B., Liu, Y. Integrating prosodic features in extractive meeting summarization. pp. 387-391 (2009).
- [10] Araly, B., Verma, R. Automated extractive single-document summarization: Beating the baselines with a new approach pp. 268-269 (2011).
- [11] Mendoza, M., Bonilla, S., Noguera, C., Cobos, C., León, E. Extractive single-document summarization based on genetic operators and guided local search. pp 4158-4169 (2014).
- [12] Yadav, J., Meena, Y.K. Use of Fuzzy Logic and Wordnet for Improving Performance of Extractive Automatic Text Summarization. pp. 2071-2077 (2016).
- [13] Nallapati, R., Zhai, F.F., Zhou, B.W. Summarunner: A recurrent neural network-based sequence model for extractive summarization of documents (2017).
- [14] Boudjehem, Djalil, and Badreddine Boudjehem. "Improved heterogeneous particle swarm optimization." *Journal of Information and Optimization Sciences* 38.3-4 : 481-499 (2017).
- [15] Begum, M., Sezer, E.A., Akcayol, M.A. Candidate sentence selection for extractive text summarization. (2020).
- [16] Darshna, P., Shah, S., Chhinkaniwala, H. Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique. pp 167- 177 (2019).
- [17] Khajeh, A., Ghasemi, M. R., & Arab, H. G. Modified particle swarm optimization with novel population initialization. *Journal of Information and Optimization Sciences*, 40(6), 1167-1179 (2019).
- [18] Rada Mihalcea and Paul Tarau, TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain. Association for Computational Linguistics. (2004).
- [19] Ajinkya Gothankar, Lavish Gupta et. al. Extractive Text and Video Summarization using TF-IDF Algorithm, *International Journal for Research in Applied Science & Engineering Technology* vol 10 pp 927-932 (2022).
- [20] Shah, E., Ahsan, M.K., Mazahir, I., Machine Learning Based Methodology for Depressive Sentiment Analysis Intelligent Technologies and Applications. (2022).

- [21] Asana Jain, Automatic Extractive Text Summarization using TF-IDF, at <https://medium.com/voice-tech-podcast/automatic-extractive-text-summarization-using-tfidf-3fc9a7b26f5> (2019).
- [22] Akash Panchal, NLP — Text Summarization using NLTK: TF-IDF Algorithm at <https://towardsdatascience.com/text-summarization-using-tfidf-e64a0644ace3> (2019).
- [23] Chawla, Indu and Sandeep Kumar Singh. "Automatic bug labeling using semantic information from LSI." 2014 Seventh International Conference on Contemporary Computing (IC3) 376-381(2014).
- [24] Alaliyat, Saleh & Yndestad, Harald & Sanfilippo, Filippo. Optimisation Of Boids Swarm Model Based on Genetic Algorithm and Particle Swarm Optimisation Algorithm (Comparative Study). (2014).
- [25] Indu Chawla, Sandeep K Singh. Automatic bug labeling using semantic information from LSI, (2014).
- [26] Heppner, H and Grenander, U., Stochastic non-linear model for coordinated bird flocks pp.233–238. (1990).
- [27] Yiyi Wang, Koji Okamura, Chapter 33 Automatic Generation of E-Learning Contents Based on Deep Learning and Natural Language Processing Techniques, Springer Science and Business Media LLC, (2020).
- [28] Roy, A., Razia, S., Parveen, N., Rao, A. S., Nayak, S. R., & Poonia, R. C. Fuzzy rule based intelligent system for user authentication based on user behaviour. *Journal of Discrete Mathematical Sciences and Cryptography*, 23(2), 409–417 (2020).
- [29] Indian Language News Text Summarization Dataset, <https://www.kaggle.com/datasets/deekoul/indian-language-summarization>, (2022).