International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

A Comprehensive Review of Marathi Text Summarization Techniques

Aarya Shah[1], Dev Patel[2], Sagar Salvi[3], Minal Sonkar[4]

[1]Department of Computer Engineering, K.J. Somaiya Institute of Technology, Sion, Mumbai, Maharashtra, India
Email: *aarya19[at]somaiya.edu*

[2]Department of Computer Engineering, K.J. Somaiya Institute of Technology, Sion, Mumbai, Maharashtra, India
Email: *dev03[at]somaiya.edu*

[3]Department of Computer Engineering, K.J. Somaiya Institute of Technology, Sion Mumbai, Maharashtra, India
Email: *sagar.salvi[at]somaiya.edu*

[4]Assistant Professor, Department of Computer Engineering, K.J. Somaiya Institute of Technology, Sion Mumbai, Maharashtra, India
Email: *minal.sonkar[at]somaiya.edu*

Abstract: ***The summary of text is the part of Natural Language Processing (NLP), which deals with condensing long text into short, meaningful and remaining important pieces of text. Text summarization in regional languages is examined specifically, with focus on their linguistic and cultural problems along with text summarization techniques. Two prominent approaches are explored: extractive summarization, and abstractive summarization which filters important sentences or phrases from the original text, and generates new phrases that contain the essence of the content. Each algorithm is evaluated, used, and applied on regional language datasets with strengths and limitations of the techniques highlighted. Because they are simple and effective for preserving semantic integrity of the source text, extractive methods are more commonly applied. In contrast, the more natural and human-like summaries produced by the abstractive methods have two major drawbacks: semantic inaccuracies and inconsistent results, to name a few, and the difficult task of dealing with various linguistic structures. Finally, this paper highlights the importance of continued research and creating such models more reliably, so we may utilize them to support text summarization in regional languages.***

Keywords: Natural Language Processing, Regional Language Text Summarization, Extractive Summarization, Abstractive Summarization, Term Frequency-Inverse Document Frequency (TF-IDF), TextRank Algorithm, Long Short Term Memory (LSTM).

1. Introduction

The quick development of multimedia technology has led to an explosion of content on many platforms, and we must create tools to condense that content for easier consumption, especially if you add language barriers. In the places where people communicate mostly in regional languages such as Marathi language, the text to text summarization techniques are being used. The current summary work has largely focussed on globally dominant languages and their regional languages like Marathi have been neglected. The overall aim of this review is to investigate the efficacy of various regional language summarization techniques, in particular, for Marathi.

While there have been advances in the techniques for summarization for globally dominant languages such as English, the regional language like Marathi has not been given much attention. This discrepancy is a result of several factors including lack of annotated datasets, as well as linguistic complexity, and relatively lower commercial incentives for these languages. This review aims to fill the gap in the understanding of regional language summarization, by providing an investigation into the efficacy of different regional language summarization strategies, in particular for Marathi.

Regional language summarization is a social and cultural necessity, as well as a technical challenge. Many rely on digital platforms for news, education and entertainment; however, millions of Marathi speakers are unable to access and process vast amounts of content due to the lack of effective summarization tools. Marathi specific summarization systems will help local communities, serve local content and help the digital age preserve the linguistic heritage.

Natural language processing (NLP), especially text summarization, has been revolutionized by deep learning. As transformer based models like BERT, GPT, T5 have gained popularity, more high-quality text summaries have been generated using them. Then, these models are pre trained on large datasets, and these are models that are absolutely amazing at understanding the context that was going on in that sentence or piece of text. But they need to be tuned with domain specialized datasets for adaptation to regional languages — Marathi in this case. In addition to such deep learning methods as seq2seq (sequence to sequence) frameworks, attention mechanisms, and hierarchical neural networks, crucial textual features are also identified and synthesized. The technique provides summary generation using these techniques that generate summaries with yield, retaining the core meaning yet accommodating the specific syntactic and semantic structures of Marathi.

The focus of this review paper is to fill the gap in summarization research for regional languages by synthesizing the existing studies in text to text summarization for low resource languages using Marathi as a focus language. Through a detailed analysis of the methodology, dataset, result, and limitation of research data of diverse regional language summarization approaches, we aim to scrutinize the

possibility of developing effective, and context sensitive summarization models for Marathi. We also present some key areas for future research that leverage emerging technologies and interdisciplinary approaches towards tackling current challenges in the region of language processing, as well as amplifying summarization capabilities for low resource languages.

2. Survey

Text summarization, one of the most important areas of work in the NLP domain. Natural language processing is dedicated to creating concise representations of significant portions of text towards meaningful content. The main purpose of summarization is usually meant to sift through the most important information from vast volumes of text to enable readers to get a quick idea about what is important. Typically, summarization techniques fall into two main categories i.e. extractive and abstractive techniques.

a) Extractive Text Summarization

Extractive summarization technique is used for summarization and the output contains important texts or lines from the source of information. This approach is aimed at discovering that portion of content which is most relevant and essential, yet provides the best viable summary of the original content, incorporating original wording and structure.

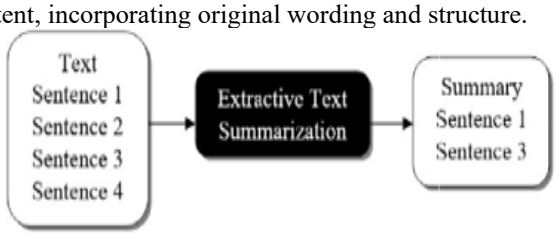

Figure 1: Extractive Text Summarization

To improve the relevance and to reduce the amount of content, TextRank algorithm is used to preprocess as well as summarize the Marathi e-news articles resulting in summarized content containing most relevant sentences for the user interests [1]. It is highlighted that the TF-IDF and Graph-based TextRank techniques in the Indian languages can be used to build extractive summarization techniques of Marathi text, which produce different outcomes when compared against each other [2].

The Graph-based TextRank algorithm generates the summary by choosing important sentences, while offering the latest Marathi e-news in a compact form with the opportunity to choose the exact word count for the summary [3]. The efficiency of extractive techniques of Indian language text summarization was compared to English language text summarization techniques out of all the text summarization methods developed for Indian languages, analyzing Precision, Recall, and F1 Score for each [4].

Automated extractive text summarization for Marathi involves producing a summarized version of a Marathi text document using only those important sentences without changing the overall meaning of the document. It makes the information more concentrated, but does not erase the main points [5]. An RNN based translation model is trained to translate text to English and a summary of the English text, and finally translating it back to Marathi. The effectiveness of the translation is then assessed using the model-based evaluation by means of BLEU and recall, for the general quality of the summaries [6].

The TextRank algorithm is used to create a word stem dictionary of selected Marathi news article's stem words and to generate a concise summary using the TextRank algorithm by extracting the relevant sentences from a news article [7]. The proposed work applies the ILrLSUMM model, an evolutionary algorithm-based technique to provide table-based summaries for both Hindi and Gujarati languages and achieves ROUGE-F-1 scores that are 34% higher than LLM for the Hindi language and 53% higher for Gujarati language [8].

The Term Frequency-Inverse Sentence Frequency (TF-ISF) model is used for summarization of Marathi documents which provides an efficient way to segment the text, tokenize Marathi documents, remove Marathi stop words and applying stemming that would help to improve the summary's relevance and accuracy [9]. LINGO Clustering designed Marathi documents into the various domains of interest and the RAND measure is reported as high as 95.83% general category documents and 93.93% of the news articles [10].

Table I: Review of Extractive Techniques

Ref No.	Key Focus	Methods / Algorithms Used	Results
[11]	Focuses on single document text summarization for marathi language.	LexRank algorithm with Gensim, a graph based technique.	A frequency based method achieved 78% precision, 72% recall, and 75% F measure while a Lex Rank algorithm achieved 78% precision, 78% recall and 75% F measure.
[12]	An effective summary must be produced with less time and less redundancy by text summarization software.	Text-rank graph based ranking model	Similarity based summarization Technique is more efficient and accurate
[13]	Summarization of marathi e-news articles using TextRank algorithm.	TextRank algorithm with Gensim library	The outcomes of text summarizing News articles using textRank algorithm are found that textRank works well with the Banking, Film Industry and other general knowledge.
[14]	Extract key points from the entire document to generate a summary.	Term frequency based algorithm, Luhn's algorithm and cosine similarity method	The cosine similarity approach is pretty faster and more accurate to generate a summary for a paragraph

[15]	Comprehensive examination of extractive text summarization techniques customized for three major Indian languages: Hindi, Bengali, and Marathi	Facial landmarks, HoG features, SVMTF-IDF, Latent Semantic Analysis	Extractive summarization is easier than abstractive summarization
[16]	Accountability of marathi text summarization techniques	Extractive summarization techniques	Reviews various extractive marathi text summarization techniques
[17]	Getting the right ranking order of sentences in the document in terms of their importance	Singular Value Decomposition and Fuzzy algorithm	The fuzzy logic algorithm was better while working on single-document summarization while SVD was better on multi-document summarization
[18]	The task involves managing documents and summarizing the Marathi text using tokens to overcome scalability issues.	TF-IDF, Cosine based document similarity	It summarizes the clusters of Marathi corpus in contrast to other past works which concentrate on the single document summarization.
[19]	They recognize condensing how text documents are into shorter and more accurate representation of meaning of the text precisely	Fuzzy Logic	Model shows that the performance has been enhanced along with enhancement in f values.
[20]	Rank the sentence according to the TF-IDF value for generating a content summary	TF-IDF	TF-IDF has precision of 0.28, recall of 0.20 and F- measure of 0.24

Thus, extractions methods TextRank, TF-IDF, LexRank show that they are effective on simple and complex Marathi texts to have more general and meaningful summaries. By selecting and post-processing the sentences, the noises involved are effectively reduced and the methods here are widely applicable for a variety of purposes including raw news articles, general and other domain-specific documents. Their progress demonstrates the increasing possibility of the Indian language summarization and opens a key direction for further improvements in the pre-processing of multilingual and regional texts.

b) Abstractive Text Summarization

Abstractive text summarization produces a brief and coherent summary for the main content while providing interpretation creating new efficient simple sentences without cutting and pasting. In this approach, its goal is to achieve a more natural and flexible model to summarize in a form closer to the form of a human summary.

Figure 2: Abstractive Text Summarization

An attention-based and stacked LSTM-based sequence-to-sequence neural network was trained and used to generate summaries for Hindi and Marathi text and the evaluation metrics ROUGE and BLEU was obtained as 0.61 and 0.638, respectively [21]. Stacked LSTM with attention and Sequence to Sequence modeling (Encoder Decoder) was employed for abstractive text summarization on the Amazon reviews having a BLEU score of 0.91 [22].

The POS taggers and named entities recognize information in Marathi text summarization of a rule-based approach based on question generation makes an abstractive intermediate summary of the input text [23]. First, the study assesses the applicability and efficiency of the proposed IndicBART model using the MahaSum dataset of Marathi news articles [24].

Built a human-annotated and expert-curated abstractive summarization data set for Telugu which is of good quality and then compared this with existing data sets in the Telugu language to test the efficacy of the set [25]. Comparing the three encoder-decoder models i.e. attention-based (BASE), multi-level (MED), and TL-based (RETRAIN) showed that by the 100k samples all models produced meaningful headlines with the RETRAIN strategy giving more substantive outcomes [26].

It adopts feature extraction through TF-IDF with stop word elimination and utilizes deep belief networks and decision trees to develop text summarization. The proposed approach yields a Mean Absolute Error (MAE) value of 2.8, precision of 95.49% and accuracy of 92.76% [27]. Presents and compares several state-of-art methods of deep learning-based abstractive text summarization on several public datasets and estimates their effectiveness by means of the ROUGE coefficients [28].

Multilingual T5-small model appears to be more accurate than IndicBART in both Hindi and Marathi, as well as in Gujarati, in essence giving a better performance on all three languages [29]. A comprehensive review of the state-of-art on summarization using machine learning focusing on feature extraction, sentence retrieval, and summary construction. The paper reviews the recent development in the categorization of extractive and abstractive text summarization methods that employ graphs, semantics, and optimization. [30].

Table II: Review of Abstractive Techniques

Ref No.	*Key Focus*	*Methods / Algorithms Used*	*Results*
[31]	Solve the problem of users to encapsulate	Wav2vec 2.0, SVM classifiers	Improved accuracy: 1.23% (detection), 10.62% (severity)

	large amount of data manually		
[32]	Combining text and image information which helps in addressing the semantics and contextual relationship by proposing Multimodality Image Text (MIT)	LSTM and Multimodality Image Text (MIT)	Proposed model achieves the ROUGE-1 score of 52.33%, ROUGE-2 score of 34.18%, and ROUGE-L score of 45.33%
[33]	Summarizing Covid-19 news to as close as human method of summary	LSTM with Deep Learning techniques	It shows that a data augmentation model performs 20% better than models without.
[34]	Capturing local and global information from the text entered by the user.	Long-Short Transformer	Introducing the transformer based model improved the quality if the summary generated significantly
[35]	Summarize the information from medical documents and extract useful information from them.	T5, BART, PEGASUS	Model PEGASUS is recognized as the better option among other models with the best ROUGE score 0.37.

In conclusion, abstractive summarization in regional languages is an interesting, yet underexplored, avenue in natural language processing. Despite its more human leaning approach to generating concise and meaningful summaries, however, the lack of research in this arena as well as results for regional languages are quite challenging. Therefore, the number of papers that refer to abstractive summarization techniques in regional languages is much less than for extractive techniques. References to studies of English text summary using abstractive methods have been added to bridge the knowledge gap and gain a more profound understanding of abstractive summarization methods.

3. Results and Discussion

In analysis of several Marathi Text summarization techniques, it is discovered that the technique is effective, advantageous and vulnerable. TF-IDF and TextRank type of extractive methods work well in retaining important sentences, but they do not create coherent summaries. However, abstractive methods including transformer based BART and mT5 achieve good readability and contextual understanding, however these models need to spend large amounts of training data and computational resources.

It is seen from the surveyed literature that hybrid approaches blending extractive and abstractive approaches are good at tradeoff between informativeness and fluency. Furthermore, there are few data sources for Marathi text summarization which hinders the development of robust models. Previous experiments conducted in the above literature survey states that extractive summaries gets more evaluation metric, e.g. ROUGE score(s), but human evaluation shows that abstractive summaries are more powerful.

4. Conclusion and Future Work

In conclusion with a review of the progress and the limitations of regional language text summarization in both extractive and abstractive fashion. Extractive approaches, although implemented relatively easily, have garnered limited research whereas such abstractive summarization can produce human-like, coherent summaries, which are yet to be explored while using regional languages. Key research missing in this domain is due to the lack of well-developed studies and information resources, as well as the inaccessibility of multiple disparate dataset, and the intricacy of language and cultural difference in regional language.

Future in this field should put greater focus on the building of robust models for (specific) regional languages with businesses taking advantage of transfer learning, cross lingual embeddings and multimodal data integration. The advancements in such models, while improving the accuracy and applicability of summarization models, will also help bridge the digital divide for non-English speaking communities. To make progress in this domain, we will need to address challenges to create the datasets, dealing with linguistic diversity and cultural specificity.

References

[1] Dhawale, A.D., Kulkarni, S., & Kumbhakarna, V.M. (2020). Automatic Pre-Processing of Marathi Text for Summarization. Regular.
[2] Manasi Chouk, Neelam Phadnis, "Text Summarization Using Extractive Techniques for Indian Language," International Journal of Computer Trends and Technology, vol. 69, no. 6, pp. 44-49, 2021
[3] Dhawale, A.D., Kulkarni, S., & Kumbhakarna, V.M. (2020). A Machine Learning Approach for Automatic Unsupervised Extractive Summarization of Marathi Text.
[4] Verma, Pradeepika and Anshul Verma. “Accountability of NLP Tools in Text Summarization for Indian Languages.” Journal of scientific research (2020): n. pag.
[5] Dhawale, A. D., Kulkarni, S. B., & Kumbhakarna, V. M. (2020). Automatic Unsupervised Extractive Summarization of Marathi Text Using Natural Language Processing. IOSR Journal of Computer Engineering (IOSR-JCE), 22(6), Ser. II, 21-25
[6] Chaudhari, A., Dole, A., & Kadam, D. (2019). Marathi text summarization using Neural Networks. International Journal for Advance Research and Development, 4, 1-3.
[7] Sankhe, S., Mahajan, M., Shinkar, B., & Patil, S. (2024). Marathi Text Summarizer. International Journal For Multidisciplinary Research.
[8] Parmar, J., Saini, N., & Dey, D. (2024). An Unsupervised Evolutionary Approach for Indian Regional Language Summarization. 2024 IEEE Congress on Evolutionary Computation (CEC)
[9] Dakulge, Umakant and S. C. Dharmadhikari. “Automated Text Summarization: A Case Study for Marathi Language.” Data mining and knowledge engineering 6 (2014): 99-105.
[10] Sahani, A., Sarang, K., Umredkar, S., & Patil, M.

(2016). Automatic Text Categorization of Marathi Language Documents.
[11] Kakde, Mrs. Kirti Pankaj and Dr. H. M. Padalikar. "Marathi Text Summarization using Extractive Technique." International Journal of Engineering and Advanced Technology (2023): n. pag.
[12] Rathod, Y.V. (2018). Extractive Text Summarization of Marathi News Articles.
[13] Dhawale, A.D., Kulkarni, S., & Kumbhakarna, V.M. (2022). Experimental Evaluation and Approach of Enhancement in Generation of Automatic Unsupervised Extractive Text Summarization of Marathi Text By Using Machine Learning Algorithm. Journal of Machine and Computing.
[14] S. Siddika and M. S. Hossen, "Automatic Text Summarization Using Term Frequency, Luhn's Heuristic, and Cosine Similarity Approaches," 2022 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET), Rajshahi, Bangladesh, 2022
[15] Basu, A., Chatterjee, A., Ghosh, R., Dasgupta, S., Roychowdhury, T., Dutta, P. K., Bhattacharya, P., & Tanwar, S. (2024). Analysis and Performance of Text Summarization Tools Applied on Indian Languages.
[16] Huici, H. D., Kairuz, H. A., Martens, H., Van Nuffelen, G., & De Kapse, R. S. (2022). Text Summarization: A Study of Marathi Language. Journal of Interdisciplinary Cycle Research, 14(7). ISSN: 0022-1945.
[17] Giri, Virat V., M. M. Math, and U. P. Kulkarni. "Marathi Extractive Text Summarization using Latent Semantic Analysis and Fuzzy Algorithms."
[18] Bafna, Prafulla B., and Jatinderkumar R. Saini. "Marathi text analysis using unsupervised learning and word cloud." International Journal of Engineering and Advanced Technology 9.3 (2020): 338-343.
[19] A. Sharaff, A. S. Khaire and D. Sharma, "Analysing Fuzzy Based Approach for Extractive Text Summarization," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 906-910, doi: 10.1109/ICCS45141.2019.9065722.
[20] Shelke, Mrs Vishakha, et al. "Text Summarization: Providing a Summary of any given input in a different language." (2022).
[21] Karmakar, R., Nirantar, K., Kurunkar, P., Hiremath, P., & Chaudhari, D.D. (2021). Indian Regional Language Abstractive Text Summarization using Attention-based LSTM Neural Network. 2021 International Conference on Intelligent Technologies (CONIT), 1-8.
[22] M. Singh and V. Yadav, "Abstractive Text Summarization Using Attention-based Stacked LSTM," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonepat, India, 2022, pp. 236-241, doi: 10.1109/CCiCT56684.2022.00052
[23] Gaikwad, D.K. (2018). Rule Based Text Summarization for Marathi Text. Journal of Global Research in Computer Sciences, 9, 19-21.
[24] Deshmukh, P., Kulkarni, N., Kulkarni, S., Manghani, K., & Joshi, R. (2024). L3Cube-MahaSum: A Comprehensive Dataset and BART Models for Abstractive Text Summarization in Marathi. Proceedings of L3Cube Labs, Pune Institute of Computer Technology, Pune, India
[25] Urlana, A., Surange, N., Baswani, P., Ravva, P., & Shrivastava, M. (2022). TeSum: Human-Generated Abstractive Summarization Corpus for Telugu. Proceedings of the 13th Conference on Language Resources and Evaluation
[26] Lal, D. M., Rayson, P., Singh, K. P., & Tiwary, U. S. (2024). Abstractive Hindi Text Summarization: A Challenge in a Low-Resource Setting. Lancaster University, United Kingdom and IIIT Allahabad, Prayagraj, India
[27] Kale, S.D., Mahalle, P.N., Kachhoria, R., Kumar, S., Chaudhari, P., & Patil, V.D. (2023). Marathi text summarization through NLP and deep learning mechanism. Journal of Autonomous Intelligence.
[28] Zhang, M., Zhou, G., Yu, W., Huang, N., & Liu, W. (2022). A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning. Computational Intelligence and Neuroscience, Volume 2022
[29] M. Sharma, G. Goyal, A. Gupta, R. Rani, A. Sharma and A. Dev, "Evaluating Multilingual Abstractive Dialogue Summarization in Indian Languages using mT5-small & IndicBART," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024, pp. 1-6, doi: 10.1109/I2CT61223.2024.10543588.
[30] P. Janjanam and C. P. Reddy, "Text Summarization: An Essential Study," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862030.
[31] S. Modi and R. Oza, "Review on Abstractive Text Summarization Techniques (ATST) for single and multi documents," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2018, pp. 1173-1176, doi: 10.1109/GUCON.2018.8674894.
[32] S. Rafi and R. Das, "Abstractive Text Summarization Using Multimodal Information," 2023 10th International Conference on Soft Computing & Machine Intelligence (ISCMI), Mexico City, Mexico, 2023, pp. 141-145, doi: 10.1109/ISCMI59957.2023.10458505.
[33] C. Limployppipat and N. Facundes, "Abstractive Text Summarization for Covid-19 News with Data Augmentation," 2022 International Conference on Digital Government Technology and Innovation (DGTi-CON), Bangkok, Thailand, 2022, pp. 56-59, doi: 10.1109/DGTi-CON53875.2022.9849194.
[34] S. Ji and B. Yang, "Abstractive Text Summarization Based on Long-Short Transformer," 2023 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 2023, pp. 0691-0700, doi: 10.1109/AIIoT58121.2023.10174260.
[35] E. Lalitha, K. Ramani, D. Shahida, E. V. S. Deepak, M. H. Bindu and D. Shaikshavali, "Text Summarization of Medical Documents using Abstractive Techniques," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 939-943, doi: 10.1109/ICAAIC56838.2023.10140885.