JOURNAL OF HARBIN ENGINEERING UNIVERSITY
ISSN: 1006-7043

Vol 44 No. 7
July 2023

Automatic text summarization for the Hindi language: Comprehensive study and analysis

Babita Verma[1], Dr. Ani Thomas[2]
[1]Research Scholar, CSVTU, Chhattisgarh, Assistant Professor, Department of Information Technology
[2]Professor, Department of Information Technology,Bhilai Institute of Technology, Durg(Chhattisgarh), India
Email:-[1]babita.verma@bitdurg.ac.in, [2]ani.thomas@bitdurg.ac.in

ABSTRACT-The term "text summarizing" refers to the process of compressing long passages of text. The goal is to create an understandable, fluid, and coherent summary, using only the main points discussed in the document. Automatic text summarization (ATS) is a common problem in machine learning and natural language processing.Since around 1958[1] the name automatic text summarization (ATS) has been widespread and an important area of research. Hindi is the official language of India. Hindi literature is increasingly being used for various purposes. Hindi is used in many places including offices, schools, and colleges. To convert the available data into more comprehensive and relevant useful information, the automated text summarization technique (ATS) is essential due to the increasing number of texts available in Hindi. Hindi is widely spoken in India and several neighboring countries, but there is no standard way to summarize the language.The various features of ATS are presented in this study to give researchers a thorough overview, including background information, methodologies, applications, building blocks,and potential forecasting research ways.

Keywords: abstractive ATS, extractive ATS, automatic text summarization (ATS), natural language processing (NLP), Hindi-ATS, Systematic Literature Review Approach (SLRA).

1 INTRODUCTION

The World Wide Web (WWW) is a large piece of network data that is growing in size every day at an amazing rate. If it is documented unnecessarily, time is wasted looking for unnecessary or additional information. It's good to finally get automated text summaries much faster[2]. The main goal of automatic text summarization is to reduce the recognition time, speed up the process of data exploration, and produce potentially relevant data. Decide whether to read the full report and summarize the data that users see online. Preparing documents can help in everyday life. Examples include news headlines, film reviews, abstracts of research papers, and book reviews [3]. Moreover, it is constantly improved when used in media business areas, for example, text mining, data retrieval, word processing, and applications, among others.The extraction process sequentially selects the essential terms from the input documents and it combines the highest-scoring terms into a summary[4].

2 LITERATURE REVIEW

Text summarization is emerging more in today's world. As very little work has been finished in the Hindi language. The connected work that has been examined is introduced in this segment. A text summarization task using an abstract model based on the thematic model of the Hindi system. They use an extraction strategy that selects meaningful phrases based on thematic patterns. The algorithm scores sentences based on the presence of a thematic root word. The expressions with the most noteworthy scores are summed up[5]. A graph-based approach is utilized for programmed text synopsis in Hindi. They used weighted graphs as the basic method and a result for precision, recall, and F measures of 79%, 69%, and 70% of the key sentences, and used semantic analysis to identify relationships to recognize[6]. Another method for summarizing Hindi text is used for multiple texts using the extraction method. Summary updates generated from fuzzy logic are used [7]. The calculated precision values were 72.62%, the

recall was 62%, and the F measure was 67.67%. A standalone Hindi text extraction algorithm is using statistical and linguistic-based sentence extraction. They checked the subject-object-verb (SOV) of the sentences by marking the POS of the words using Hindi Word-Net. They also used a genetic approach to create summaries designed to maximize interesting topics and minimize confusion [8]. Gupta M. et al. have proposed another model for Automatic Text summarization with the help of a principle-based method [9]. An automatic graph-based semantic text summarization task is presented using particle set optimization in Hindi. Use linguistic, semantic, and syntactic tools to draw and create meaningful graphs. Performance measurements reported are 60% recall, 42.86% accuracy, and 50.01% F measurement [10]. The source report's rundown is delivered by making a rich semantic graph for the principal record and a short time later decreasing the made semantic graph to a more detached outline, and from this reduced diagram an abstractive framework is made [11]. Single document summarization involvesthe extraction strategy for Hindi message, which utilizes measurable and phonetic elements, Average TF-ISF, sentence length, sentence position, mathematical information, sentence-to-sentence likeness, title highlight, SOV capability, subject similitude and furthermore utilizes hereditary calculation [12]. Text summarization utilizing the extractive strategy utilizes three significant calculations, fluffy classifier, brain organization, and worldwide inquiry improvement (GSO). The proposed approach accomplished a normal review pace of 0.88 and a normal accuracy pace of 0.90 for a pressured pace of 20% [13]. A summary is generated on the basis of some feature extraction such as inverted commas, sentence length, sentence overall position, presence of keywords, numeric data in sentences, etc. [14].

3. ATS APPROACHES

Here are three main abstraction methods: extraction, abstraction, and hybrid.

3.1 Extractive Text Summarization

Extractive summarization techniques work by extracting meaningful expressions from the input text and transcribing them verbatim as part of the summary[15-17]. Each summary extraction technique uses the following basic operations shown in Figure 1.

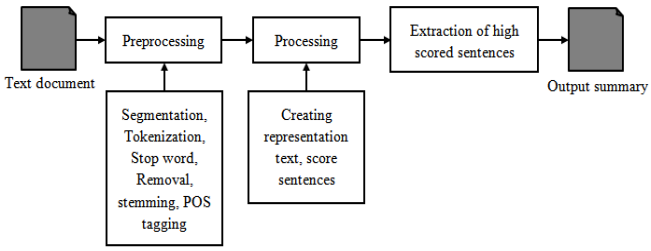

Figure1. Basic steps involved in extractive text summarization

3.2 Abstractive Text Summarization

Random summarization methods use the most efficient natural language processing techniques to understand the text and generate new syntactic messages, rather than choosing the best option available to the agent to summarize.[16-17]Each summary abstractive technique uses the following basic operations shown in Figure 2.

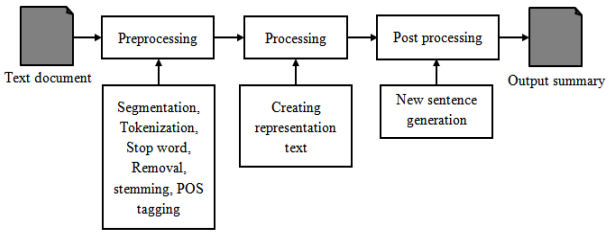

Figure2. Steps involved in abstractive text summarization

3.3 Hybrid Text Summarization

The hybrid system combines the principles of both extraction and desorption. [15-16] The steps involved in the hybrid text summarization system are shown in Fig. 3.

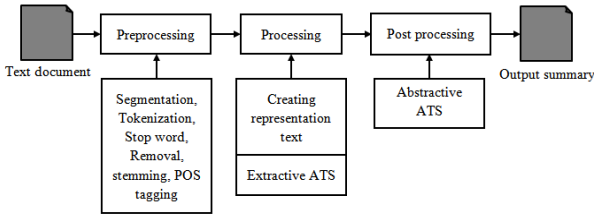

Figure 3. Steps involved in the hybrid text summarization system

The various methods used for text summarization under the basic text summarization approach are shown in figure-4.

4. RESEARCH METHODOLOGY

This paper's research methodology is a systematic literature review approach(SLRA). This study's primary goal is to present the most recent work of several research scholars and construct a model that produces a summary of the text from various application areas—reviewed around 50 research papers to increase the impact of our research in terms of systematic review selection and outcomes. A screening protocol has been introduced on the basis of the research questions.

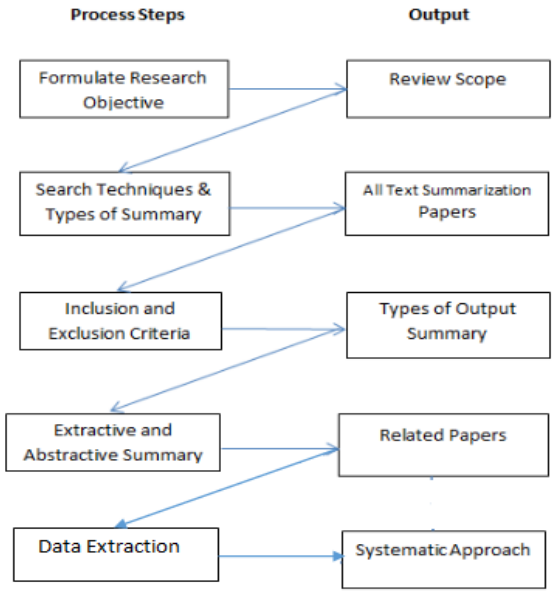

Fig. 4. An SLRA Process Model

4.1 Research Objectives

The first step of the SLRA is to formulate research objectives. The main identified objectives of this research study are as follows:

1. The major goal is to pinpoint the gaps in the literature and difficulties that various text summarizing techniques confront.
2. Automatic summarization is being done for popular languages like English. Less research work has been done on Hindi languages.
3. Few works are carried out in the field of extractive summarization for Hindi Language and to an extent possible for abstractive Summarization too.

4.2 Search Techniques

The second stage of SLRA entails gathering published research articles and approaches for text summarising.The following digital libraries are where we found the publications we used as the basis of our research: Google Scholar, IEEE, Springer, and Science Direct as presented in Table 1.

Sources	Search String	Context
Google Scholar, IEEE Explore, Science Direct, Springer showed in fig 6.	"Automatic Text Summarization Techniques for Hindi Document" "Text Summarization based machine learning OR deep learning" and "Text Summarization techniques using NLP" or "Automatic Hindi Text Summarization"	Text Summarization

Table 1. Terms and keywords used in the search

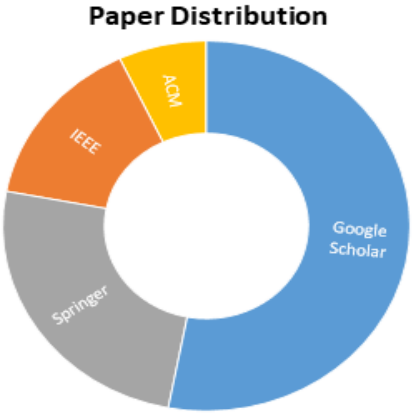

Figure 6. Distribution of relevant search papers

4.3 Inclusion Criteria

In an SLRA, we emphasize high-quality articles that outline several approaches to text summary utilizing various techniques and methods. The further inclusion standards are specified:

- A study on text summarization in many languages is discovered.

- Several approaches using various dataset types are presented in research articles.
- Several Text Summarization methods, including ML, NLP, and DL, are presented in research articles.

4.4 Exclusion Criteria

In an SLRA, Only papers that are relevant to our topic and search string are included, whereas certain papers that exist in several sources are excluded. Also found some papers which are not related to our search string. The further exclusion standards are described:

- Unrelated research articles on text summarization.
- Text classification techniques and strategies are the subjects of research publications.

4.5 Search Collection

The publications in our search must be sorted based on actual relevance because not all of them are relevant. The identical paper was only taken into account once in accordance with our search order when it appeared in multiple sources. Papers that weren't fit for the research field were excluded from consideration in the first round of screening based on their titles. For instance, sometimes the articles returned by our search string are either unrelated to our field of research or have a distinct meaning. We reviewed the abstracts of the papers that were chosen during the first step of screening in this phase.

4.6 Data Extraction

The Google Scholars database contains a total of 11,490 citation summary publications from 2000 to 2021 (to date). The number of baseline summaries increases every year. Figure 1 shows the annual trend of the aspect.

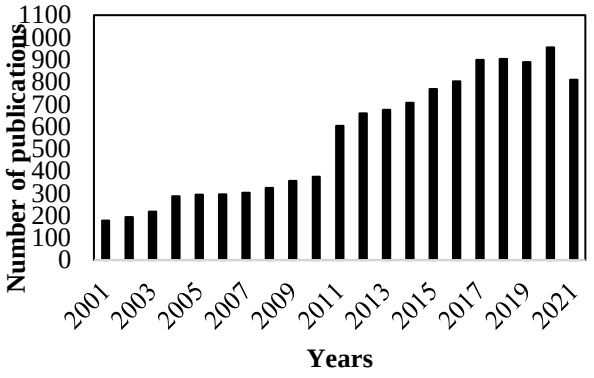

Figure 7 Year-wise publications in the research area of text summarization

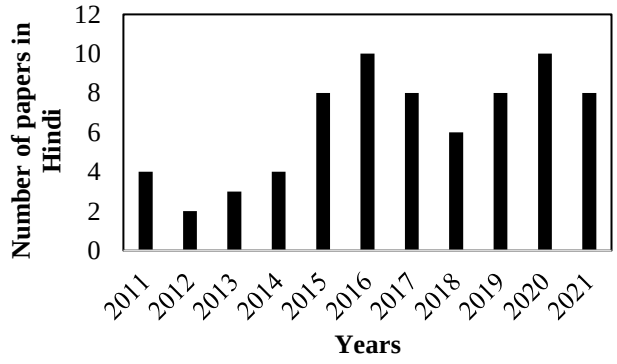

Figure8. Year-wise publications in the research area of Hindi text summarization

The Google Scholars database was used to review previous research work to summarize the literature. Data from the past 11 years were examined to examine publication trends.In India, Hindi is the most common official language. According to the Indian government survey, the percentage of Hindi speakers is 41.03 and 53.6% respectively.Summary Among all books published in the literature (in different languages), the number published to summarize Hindi literature is very low, that is 11490 with only 71 essays in Hindi. To summarize the Hindi literature shown in Figure 3, a lot of research is needed.

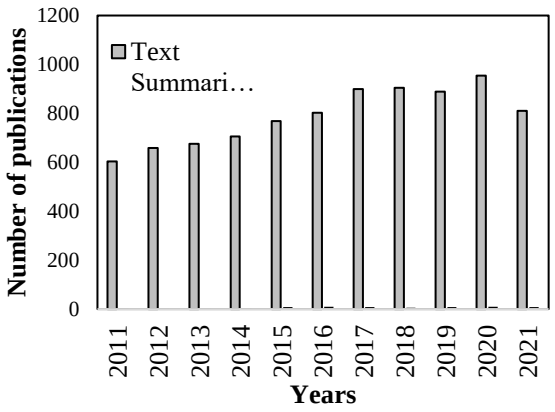

Figure 9. Comparative studies of year-wise publications in text summarization and Hindi text summarization

5 COMPARISON OF VARIOUS TEXT SUMMARIZATIONAPPROACHES

The reading compares the formulation techniques that are beyond in formulation types that were made during the literature review process in Table 2.

Table 2 Comparison of summarization methods

Details	**Subtype**	**Concept**	**Advantages**	**Disadvantages**	**Application/Work Done**
Number of input document	Single document	Can acknowledge an information record.	They concentre on load overhead reduction	Summing up archives with a few related topics isn't inconceivable	H. P Edmundson used title of the word, cue phase, key method, position method – surface level approach[40]
	Multi-document	It acknowledges a few information records	Different records of the same subject can be summed up to a single archive	Real-time implementation is difficult for a designer	SUMM ONS Designed by Columbia university [38]
Language	Mono-Lingual	Only input with a specific language can be accepted and the output rely on the language	Only one language should be with work	Cannot handle different types of language	GistSumm [38]
	Multi-Lingual	It accepts documents in multiple languages	It also deals with thekinds of languages	It is difficult to insist	SUMMARIST (English, Japanese,Spanish) [39]
Content	Generic	General summary regardless of user type. Knowledge is at the same level of importance	Any user can make use of it	It alwaysprovides the view of author,nouser-specific	SUMMARIST[40]

	Query based	The user only needs to specify the content of the source document in query mode and the method of retrieving the information	Specific information could be searched. It may reflect the user's interest	It is based on who uses it. It cannot be used by any type of user	Mitre's WebSumm[40]
Details	Indicative	It essentially presents the primary thought of the article to the client. It very well may be utilized to rapidly decide if text is intelligible	It likewise urges the clients to peruse the compulsory report in profundity. Utilized for speedy grouping and simpler to make	It does not provide detailed information	The back of a movie or novel package will show 5-10% of the information
	Informative	It summarizes the main text	It is as like a substitution for the main document	It does not provide a quick overview of the information	Length 20 to 30% SumUM [33]
Limitation	Genre specific	Only accept special types of documents as input.	It is used to overcome the problems of summarizing the heterogeneous document	It has the kind of limitation templates of the text	News blaster
	Domain dependent	Summarize texts that can define their content in a specific place	They are aware of which domain they are dependent	They are limited to the subject of the document provided	TRESTLE [38]
Type of generated summary	Abstractive	The most common way of decreasing the text of a report to a significant sound outline	Good compression ratio andmuch-reduced text and separately related summary	Computation is difficult	SUMMRIST [32]
	Extractive	It consists of a selection of key phrases from the original text based on statistical diagrams	Easy to compute the text	It rather experiences irregularities, absence of equilibrium, brings about extensive outline	Summit applet,designed by Surrey University [38]

5. CONCLUSION AND FUTURE SCOPE

It has been seen that summarization of text always plays a significant role in saving users' time and resources as a result of the large amounts of data available. Text summarization is undoubtedly a crucial modern-day tool. These techniques, both individually and collectively provide several summaries. Their precision, better, and more concise scores can be determined by comparing different summaries. The ROGUE symbol is used more often than this. Similarly, the IDF sign was also used. The summaries produced using these techniques are not always accurate. It can also be unrelated to the original content. As a result, research on this subject is ongoing, and many studies have been done. In order to produce more accurate summaries in the future, the models presented can be updated. So, there's a probability that summarized the systems mentioned here will undergo changes in the following years to produce a summary that is more accurate.The two summarizing strategies used in NLP are extractive and abstractive methods. Extractive summarization methods are more prevalent in this field, but the abstractive summary technique can producea summary that is more accurate, comprehensive, and relevant to the original document. Very less research work has been done in abstractive summary generation for the Indian language. The readers of this paper will have a thorough understanding of the insufficiency of research in the area of Automatic summarization of Indian languages. Furthermore, it will provide us with the capability and assurance to precede with NLP-based summarization techniques.

Availability of Data & Material

References given in the paper

Abbreviations

ATS :Automatic text summarization

NLP : natural language processing

WWW :World Wide Web

TF IDF :term frequency-inverse document frequency

References

[1] Hans Peter Luhn. 1958. The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 2, 159–165.

[2] Dipanjan Das and Andre F. T. Martins. 2007. A survey on automatic text summarization. Lit. Survey Lang. Stat. 4, 192–195.

[3] EhsanShareghi and Leila Sharif Hassanabadi. 2008. Text summarization with harmony search algorithm-based sentence extraction. Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology.ACM.226–231.

[4] K. Sankar and L. Sobha. 2009. An approach to text summarization. Proceedings of the 3rd International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies. ACL.53–60.

[5] Kumar, K. V., &Yadav, D. (2015). An improvised extractive approach to Hindi text summarization. In Information Systems Design and Intelligent Applications (pp. 291-300). Springer, New Delhi.

[6] Kumar, K. V., Yadav, D., & Sharma, A. (2015). Graph-based technique for hindi text summarization. In Information Systems Design and Intelligent Applications (pp. 301-310). Springer, New Delhi

[7] Gulati, A. N., &Sawarkar, S. D. (2017, January). A novel technique for multi-document Hindi text summarization. In 2017 International Conference on Nascent Technologies in Engineering (ICNTE) (pp. 1-6). IEEE.

[8] ChetanaThaokar, Latesh Malik, “Test Model for Summarizing Hindi Text using Extraction Method”, Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).

[9] Gupta, M., &Garg, N. K. (2016, September). Text Summarization of Hindi Documents using Rule Based Approach. In 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE) (pp. 366-370). IEEE.

[10] Dalal, V., & Malik, L. (2017, March). Semantic Graph Based Automatic Text Summarization for Hindi Documents Using Particle Swarm Optimization. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 284-289).Springer, Cham.

[11] SubramaniamManjula, DalalVipul, " Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method.", IRJET, 2015.

[12] DalalVipul, ShelarYogita, "A Survey of Various Methods for Text Summarization", International Journal of Engineering Research and Development, Vol. 11, Issue 03 2015, PP.57-59.

[13] AnithaJ. , Prof. Prasad Reddy P. V. G. D., Prasad Babu M. S., "An Approach for summarizing Hindi Text through a Hybrid Fuzzy Neural Network Algorithm", Journal of Information and Knowledge Management, 2014,Vol. 13, No. 4

[14] Nikita Desai, Prachi Shah, Automatic Text Summarization Using Supervised Machine Learning Technique For Hindi Langauge, International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

[15] Som Gupta, S.K Gupta, Abstractive Summarization: An Overview of the State of the Art, Expert Systems With Applications (2018), DOI: https://doi.org/10.1016/j.eswa.2018.12.011

[16] El-Kassas, W.S., Salama, C.R., Rafea, A.A., Mohamed, H.K., Automatic Text Summarization: A Comprehensive Survey, Expert Systems with Applications (2020), DOI: https://doi.org/10.1016/ j.eswa.2020.113679

[17] VipulDalal, Dr.Latesh Malik Automatic Summarization for Hindi Text Documents using Bio-inspired Computing, International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 4, April 2017 DOI:10.17148/IJARCCE.2017.64130