Review of Text Summarization in Indian Regional Languages

Surendrabikram Thapa, Surabhi Adhikari, and Sushruti Mishra

Abstract Propelled by the advancements in the field of natural language processing, generating summaries of long texts using various NLP tools and techniques has always been a subject of great interest for scientists all over the world. Data is ubiquitous, and a large amount of data is processed every second in the digital space. For these reasons also, there is need of machine learning algorithms that can automatically shorten longer texts or documents and provide the accurate and meaningful summaries. These summaries find their applications in a wide variety of fields like medicine, market review, business analytics, etc. In other languages like English, an ample amount of study and research can be observed. However, in context of the Indian regional languages, the research in text summarization is very limited and is still in the infancy state. This paper tries to explain the research and works that have been performed in the field of text and document summarization in Indian regional languages.

AQ1

Keywords Text summarization · Extractive summarization · Abstractive summarization · Indian regional languages

1 Introduction

Natural language processing (NLP) in particular deals with programming the computers to parse, process and perform analysis of huge amount of human language data that surrounds us. Today, the world is so centralized on computers, and for that

S. Thapa (✉) · S. Adhikari · S. Mishra
Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India
e-mail: surendrabikram_bt2k17@dtu.ac.in

S. Adhikari
e-mail: surabhi_bt2k18@dtu.ac.in

S. Mishra
e-mail: mishrasushruti99@gmail.com

A. Abraham et al. (eds.), *Proceedings of 3rd International Conference on Computing Informatics and Networks*, Lecture Notes in Networks and Systems 167,
https://doi.org/10.1007/978-981-15-9712-1_3

reason, natural language processing is widely used these days [1]. Text summarization is a sub-domain of NLP which deals with extracting or collecting the important information from the given text or document and gives concise information regarding the text or document in the form of brief summary. Automated summarizers reduce reading time, and, in many cases, these summarizers can provide unbiased summaries than that of human beings. Text summarization has a lot of applications like customer review summary, online news article summary, a summary of the minutes of the meeting, automated research abstract, etc. Automatic text summarizers in languages such as English were in existence from 1950, but the advancement in summarization has seen a rapid pace in the last two decades. Indian regional languages have however seen good development in recent 10 years. In India, there is no one single language that is used across India as an official language. There are more than 22 official languages in India, and each of them is used for official purposes. So, our concern should not be getting too focused on a single language, but the summarization techniques in every language should also be explored. There cannot be a single system that can generalize the summarization process of all Indian regional languages. This is because all languages have their own linguistic features, and hence, each language should be dealt with independently. But works and study in the field of such text summarization in Indian regional languages are very less and are still in infancy state.

Based on the techniques used, text summarization techniques can be broadly divided into two major categories, namely extractive summarization and abstractive summarization. Extractive method of text summarization selects phrases and sentences from the source documents or text and includes such information extracted in the newly generated summary. The summary is based on key features in the text. For finding the phrases and sentences required, extractive methods make use of statistical features like sentence position, proper noun, numerical data, topic frequency, topic token frequency, normalized sentence length, cluster frequency, etc. Mostly, extractive summarization techniques make use of three tasks viz. tokenization of the text, calculating the word scores mainly TF/IDF scores, calculation of scores of the sentences based on such word scores and selection of summary comprising highest scores. In this way, the extractive summarization method, in short, extracts the most relevant information from the text document and includes them in summary. Since the phrases or sentences of generated summaries are directly extracted from the given document or text, the summaries are sometimes not meaningful and complete. However, this method is preferred because of its easy implementation. The abstractive text summarization technique, on the other hand, generates a meaningful summary, and the summary is more human-like. This method generates entirely new phrases and sentences to provide a summary of the source text. The system generates new phrases by rephrasing instead of simply extracting phrases or even using the words that are not present in the source text or document. Since the abstracted summary may contain phrases or sentences that are not in the source document or text, this is a more challenging approach. For a good abstractive summary, the built model should be given the independent ability to understand the given document, and the model will output its understanding in its own words [2]. Mostly deep learning models are

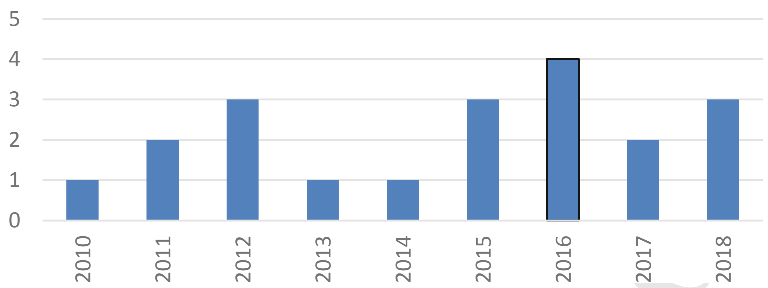

Fig. 1 Distribution of the number of papers discussed and year of publishing

used to generate an abstractive summary. It is much harder than extractive techniques, but the human-like summary outweighs its drawbacks in implementation.

In Indian regional languages, the extractive methods are very widely used. Research in abstractive methods, on the other hand, is very nominal and needs to be explored more extensively. In this paper, we present the summarization techniques that are used in various Indian regional languages in great detail. The paper describes the methods, different works carried out in various Indian regional languages and comparative analysis of different methodologies used in text summarization of Indian languages in the past 10 years, and the distribution of number of papers and year of publishing is as shown in Fig 1.

2 Related Works

With more than 122 languages and 22 designated official languages, India has a huge diversity in languages [2]. The need for text summarization in regional languages is much needed as they are used extensively in official works. A lot of effort has been made by a lot of researchers to effectively summarize the texts written in Indian regional languages. In this paper, we majorly discuss the text summarization techniques used in seven major languages used in India viz. Hindi, Bengali, Telugu, Marathi, Tamil, Urdu and Punjabi. The techniques that were used in the past 10 years for the text summarization in Indian regional languages have been discussed in this paper. Figure 2 shows the distribution of the number of papers that we have studied and the languages.

Generally, the approach followed by each of the summarizing technique can be represented by flow chart diagram as shown in Fig 3. The summarization usually starts with input of text which is followed by preprocessing steps done to it. After preprocessing of the text, there is iterative calculation based on the model of the summarizer (mostly calculations are based on word vector, word–sentence relationship and graph model). The process is followed by final step of summary generation after doing the analysis of processes done before.

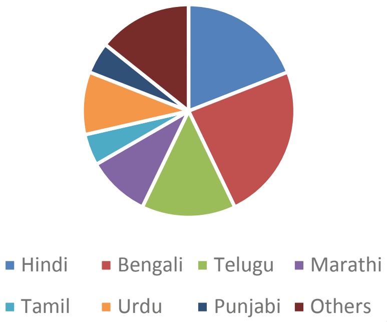

Fig. 2 Distribution of the language and the number of papers discussed

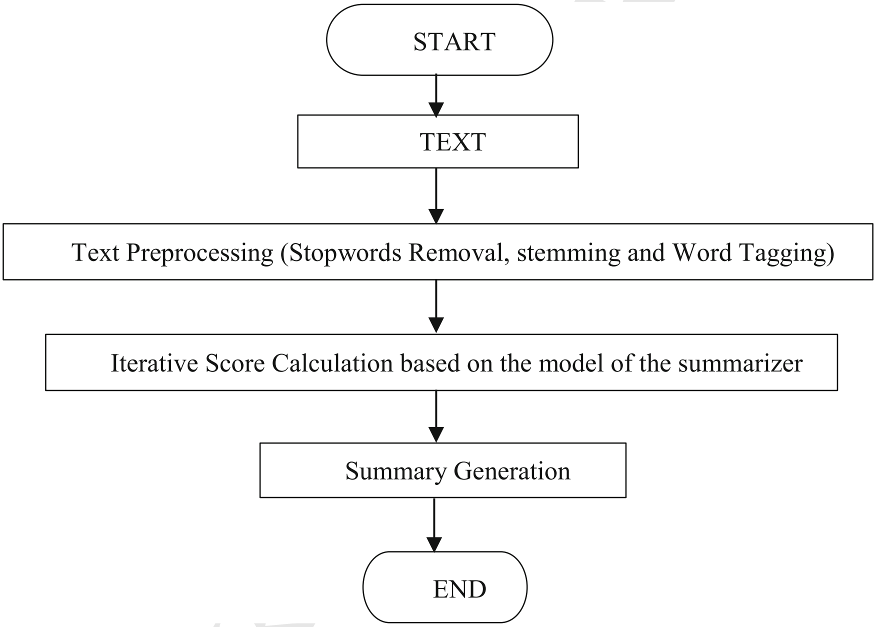
START
TEXT
Text Preprocessing (Stopwords Removal, stemming and Word Tagging)
Iterative Score Calculation based on the model of the summarizer
Summary Generation
END

Fig. 3 Flow chart for working of text summarizer in Indian regional languages

2.1 Hindi

Thaokar et al. [3] discussed on how we can carry out summarization of given text or documents in Hindi language using the sentence extraction method. The technique used six statistical features along with two linguistic features. It used Hindi WordNet for tagging of the appropriate part of speech of the word to check subject–object–verb of the given sentences. The result was optimized using a genetic algorithm. Gupta et al. [4] proposed the rule-based technique for the summary generation of the

Hindi text documents based upon the linguistic rules. Five features for each sentence were discovered, and the proposed methodology was then subject to testing on thirty different documents that belonged to various domains. Input size was decreased to 60–70% with an accuracy of 96% when testing was done on given thirty given documents for generation of the summary of provided Hindi text. Kumar et al. [5] proposed the system based on an extractive approach that selected the important and meaningful sentences from the text based on some thematic approach. The system relies on the scoring based on the occurrence of root words, and the sentences with the highest scores were then included in the generated summary. Subramaniam et al. [6] in 2015 have proposed an abstractive method for the generation of summaries.

2.2 Bengali

The extractive method that is used in conjunction with the approach proposed by Abujar et al. [7] and a set of Bengali text analysis rules derived from the heuristics was able to give summaries nearly equivalent to human-generated. In the approach proposed by the authors [7], the importance of sentences and phrases was identified based on word scoring, sentence scoring and graph scoring to find the appropriate texts that need to be included in the summary.

Sarkar [8] discussed an effective extractive summarizer that gives a summary using sentence length feature along with TF/IDF weights (TF: term frequency and IDF: inverse document frequency). The evaluation results in the paper showed that the methodology proposed had better performance in terms of different performance measures than given three systems compared to the paper.

Akter et al. [9] proposed a method that extracts important sentences or phrases from the text document. Word score in the system was calculated by TF/IDF. After calculation of word scores, sentence scores were calculated. The scores of words that constituted the sentence were added, and value for position of sentence was also given while calculating sentence scores. Finally, the K-means clustering algorithm was used for the generation of required summary. Das et al. [10] proposed another topic-based opinion summarizer for Bengali language. Features were extracted in forms like syntactic and lexico-syntactic features. Aggregation of such topic-sentiment was then done using a clustering algorithm (K-means).

2.3 Telugu

Telugu is a Dravidian language and is mostly spoken by the people residing in Indian states of Telangana, Andhra Pradesh and Union Territory of Puducherry. Reddy et al. [11] proposed an extractive summarization technique that summarizes the articles in the Telugu language by using key features such as sentence's order of appearance in the given document, sentence similarity with title, word-frequency and centrality of the sentence. The sentences were then ranked by calculating scores for each sentence by taking all given features into consideration. Naidu et al. [12] have also suggested a summarization technique that summarizes text with automatic keyword extraction from the dataset used in Telugu e-newspapers. In their described technique, the researchers have used human evaluation to train the system for seeking the key phrases or keywords that have maximum probability of inclusion in summary. They were able to get great accuracy when tested with different datasets with this method. Similarly, Kallimani et al. [13] have proposed the abstractive method for summary generation of text documents in the Telugu language.

2.4 Marathi

Giri et al. [14] discussed the extractive summarization in the Marathi language by extracting relevant sentences using the application of statistical features as well as features that depended on Marathi language. Similarly, the works of Sarwadnya et al. [15] discussed the text or document summarization technique that used a graph-based model that also used the extractive approach of text summarization in the Marathi language.

2.5 Tamil

The Tamil language is another Dravidian language that is spoken predominantly in Singapore, Sri Lanka, Tamil Nadu and Puducherry. Among the very few works done in Tamil language, Priyadharshan et al. [16] have proposed the method which can automatically summarize Tamil online sports news articles using natural language processing and a generic stochastic ANN. The feature matrix was created with various linguistic features to enhance accuracy.

2.6 Urdu

Burney et al. [17] in 2012 had designed and developed an add-in for MS word which summarized text in Urdu language. The approach that used a statistical method of sentence weight algorithm was able to summarize the Urdu text to the accuracy of over 80% when one human verifier was used. Humayoun et al. [18] have discussed the effect of preprocessing setting in the accuracy of summarization of text in the Urdu language which is one of the great works done in the field of NLP in the Urdu language.

2.7 Punjabi

Punjabi is another widely spoken language in India. Gupta et al. [19] have described the automatic summarizer for Punjabi text for summarizing the news articles in the Punjabi language. In their proposed method, the score of the sentences was calculated by making use of a feature-weight equation, and the sentences with highest rankings were then arranged to get a summary of the news articles.

2.8 Other Regional Languages

NLP and the works in summarization of text are still in the premature stage in other regional languages. In the Kannada language, Geetha et al. [20] have discussed extractive text summarization using latent semantic analysis. Similarly, Kabeer et al. [21] proposed the text summarizer for Malayalam text documents in 2014. The paper implemented a statistical technique for calculating the sentence scores and made use of a semantic graph-based method for summary generation. Also, Krishnaprasad et al. [22] have described a similar extractive approach for summarization of text in the Malayalam language. The works in other languages such as Assamese, Nepali, Santali and Dogri are very limited. Works in semantic analysis have been done, but the work in text summarization has not yet been explored.

3 Discussion

From the above works, we have seen that researchers who are doing their research in text summarization of Indian regional languages extensively use extractive text summarization techniques. We have also seen that an ample amount of work has been done in the Hindi language than any other Indian regional language. Figure 4 gives the picture of distribution of the number of papers that we have studied and the methods used in those papers. Despite some inaccuracy extractive models pose, researchers have found their way to make summaries of long text more accurate by various methods. The extractive summarization process that involves the conventional process of primarily ranking sentences with scores and including the sentences with highest scores can be improved with various techniques such as improvement in preprocessing and sentence scoring methods.

Krishnaprasad et al. [22] have suggested that our summarizers can generate even great quality summaries if we can improve the sentence scoring stage. Sarkar [8] has also suggested that the system performance may be further improved by improving stemming process including more number of effective features. Similarly, the relevancy of sentences should also be measured to get better summary. Akter et al. [9] in their work suggested that the relevancy of sentences can be measured by using

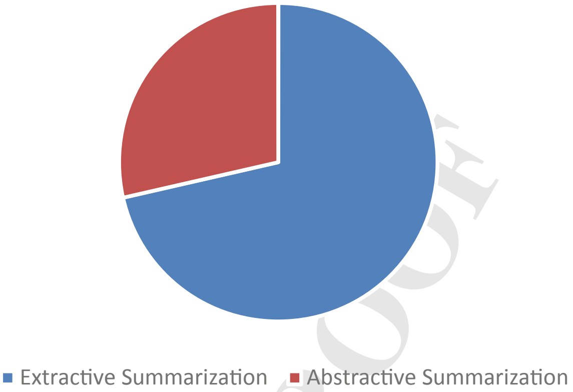

Fig. 4 Distribution of number of papers with summarization techniques used

syntactic and similarity in the future to get good accuracy. Apart from this, a lot of time and work must be done in the addition of more features to get more accurate summaries. Reddy et al. [11] who proposed their works in Telugu text summarization also suggested that with addition of more features like cue phases, the existence of some punctuation marks, day-month names, numeric, literals, etc., to the existing methodologies can boost up the quality of summary generated by the system.

The efforts have been made by authors to attain maximum accuracy with the extractive text summarization process. Sometimes, the sentences that are extremely important and are extremely important for inculcating in final generated summary are long sentences. This makes generated summaries lengthier, and we might also be missing some other relevant information that might be present in other shorter sentences when our model includes such long sentences by keeping such shorter sentences aside. Kabeer et al. [21] in their paper discuss the need for abstractive summarization processes to deal with such problems. Sunitha et al. [23] in their paper have presented experimental works that have been performed by using the abstractive summarization techniques in Indian regional language. They have also put an emphasis on promoting method of abstractive summarization in Indian regional language. With rapid advancement in deep learning and advanced researches in neural networks and various deep models, abstractive summarization can serve as a very great technique for summarizing text documents. Also, generative models can prove to be a great summarizer for text documents. Liu et al. [24] have proposed a generative adversarial network for an abstractive summary generation. Their work was concentrated on the English language but the same can also be done in Indian regional languages also. Similarly, the use of transfer learning can also facilitate the task of summarization using abstractive methods. Also, various scoring methods which provide a better method for evaluating the quality of the summary can be explored.

4 Conclusion and Future Works

The enormous amount of data that surrounds us needs summarization, and we are mostly surrounded by text that is in our regional languages. So, text summarization in regional languages should be of great priority of research. The study of above discussed paper also roughly presents that we should shift more towards abstractive summarization so that the machines can generate human-like summaries. The irrelevant information in extracted summary sometimes might lead us to confusion and that can be solved slowly with techniques that will be available with the advent of newer technologies. Also, very less works have been done in regional languages other than the abovementioned. The work should be taken with high priority especially by the native speakers of those languages. The paper has clearly explained that we lack in the field of abstractive summary and future works should be towards making summarizers for an abstractive summary generation. Similarly, further works should be done in addressing the complex morphological variations in Indian regional languages.

References

1. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12:2493–2537
2. Anthes G (2010) Automated translation of Indian languages. Commun ACM 53:24–26
3. Thaokar C, Malik L (2013) Test model for summarizing hindi text using extraction method. In: 2013 IEEE conference on information and communication technologies. IEEE, pp 1138–1143
4. Gupta M, Garg NK Text summarization of Hindi documents using rule based approach. In: 2016 international conference on micro-electronics and telecommunication engineering (ICMETE). IEEE, pp 366–370
5. Kumar KV, Yadav D (2015) An improvised extractive approach to hindi text summarization. In: Information systems design and intelligent applications, pp 291–300
6. Subramaniam M, Dalal V (2015) Test model for rich semantic graph representation for Hindi text using abstractive method. Int Res J Eng Technol 2
7. Abujar S, Hasan M, Shahin M, Hossain SA (2017) A heuristic approach of text summarization for Bengali documentation. In: 2017 8th international conference on computing, communication and networking technologies (ICCCNT). IEEE, pp 1–8
8. Sarkar K (2012) An approach to summarizing Bengali news documents. In: proceedings of the international conference on advances in computing, communications and informatics, pp 857–862
9. Akter S, Asa AS, Uddin MP, Hossain MD, Roy SK, Afjal MI (2017) An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm. In: 2017 IEEE international conference on imaging, vision and pattern recognition (icIVPR). IEEE, pp 1–6
10. Das A, Bandyopadhyay S (2010) Opinion summarization in Bengali: a theme network model. In: 2010 IEEE second international conference on social computing. IEEE, pp 675–682
11. Reddy PV, Vardhan BV, Govardhan A (2011) Corpus based extractive document summarization for Indic script. In: 2011 international conference on Asian language processing. IEEE, pp 154–157
12. Naidu R, Bharti SK, Babu KS, Mohapatra RK (2018) Text summarization with automatic keyword extraction in telugu e-newspapers. Smart computing and informatics. Springer, pp 555–564

13. Kallimani JS, Srinivasa K (2011) Information extraction by an abstractive text summarization for an Indian regional language. In: 2011 7th international conference on natural language processing and knowledge engineering. IEEE, pp 319–322
14. Giri VV, Math M, Kulkarni U (2016) A survey of automatic text summarization system for different regional language in India. Bonfring Int J Softw Eng Soft Comput 6:52–57
15. Sarwadnya VV, Sonawane SS (2018) Marathi extractive text summarizer using graph based model. In: 2018 fourth international conference on computing communication control and automation (ICCUBEA). IEEE, pp 1–6
16. Priyadharshan T, Sumathipala S (2018) Text summarization for Tamil online sports news using NLP. In: 2018 3rd international conference on information technology research (ICITR). IEEE, pp 1–5
17. Burney A, Sami B, Mahmood N, Abbas Z, Rizwan K (2012) Urdu text summarizer using sentence weight algorithm for word processors. Int J Comput Appl 46:38–43
18. Humayoun M, Yu H (2016) Analyzing pre-processing settings for Urdu single-document extractive summarization. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), pp 3686–3693
19. Gupta V, Lehal GS (2012) Automatic Punjabi text extractive summarization system. In: Proceedings of COLING 2012: demonstration papers, pp 191–198
20. Geetha JK, Deepamala N (2015) Kannada text summarization using latent semantic analysis. In: 2015 international conference on advances in computing, communications and informatics (ICACCI). IEEE, pp 1508–1512
21. Kabeer R, Idicula SM (2014) Text summarization for Malayalam documents—an experience. In: 2014 international conference on data science and engineering (ICDSE). IEEE, pp 145–150
22. Krishnaprasad P, Sooryanarayanan A, Ramanujan A Malayalam text summarization: an extractive approach. In: 2016 international conference on next generation intelligent systems (ICNGIS). IEEE, pp 1–4
23. Sunitha C, Jaya A, Ganesh A (2016) A study on abstractive summarization techniques in Indian languages. Proc Comput Sci 87:25–31
24. Liu L, Lu Y, Yang M, Qu Q, Zhu J, Li H (2018) Generative adversarial network for abstractive text summarization. In: Thirty-second AAAI conference on artificial intelligence