International Journal of Advanced Research in Engineering and Technology (IJARET)
Volume 12, Issue 1, January 2021, pp. 530-538, Article ID: IJARET_12_01_049
Available online at http://iaeme.com/Home/issue/IJARET?Volume=12&Issue=1
Journal Impact Factor (2020): 10.9475 (Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6480 and ISSN Online: 0976-6499
DOI: 10.34218/IJARET.12.1.2021.049

TEXT SUMMARIZATION FOR INDIAN LANGUAGES: A SURVEY

Kishore Kumar Mamidala
Associate Professor, Department of Computer Science and Engineering
Vivekananda Institute of Technology and Science, Karimnagar, India

Suresh Kumar Sanampudi
Assistant Professor and Head of Department of Information Technology
Jawaharlal Nehru Technological University Hyderabad, Telangana, India

ABSTRACT

With the increasing amount of huge data availability on the internet, the need for automatic text summarization has emerged in the recent past. Text summarization methods generate summaries of the relevant information from original content. Text summarization methods are two types abstractive and extractive. For English, numerous text summarization techniques exist in the literature. But for Indian languages, there are only a few techniques developed. This paper presents a survey and analysis of text summarization methods developed for Indian languages—the challenges involved in summarizing Indian language documents. Merits and demerits of the existing techniques are listed. This paper also investigates which method is ideal for summarizing documents in Indian languages.

Keywords: Natural Language Processing; Text Summarization; Extractive summarization; Statistical Methods; Machine Learning.

Cite this Article: Kishore Kumar Mamidala and Suresh Kumar Sanampudi, Text Summarization for Indian Languages: A Survey, *International Journal of Advanced Research in Engineering and Technology,* 12(1), 2021, pp. 530-538.
http://iaeme.com/Home/issue/IJARET?Volume=12&Issue=1

1. INTRODUCTION

Text Summarization is a method of extracting or deriving the abstract of the original information [2]. In Mani and Maybury [3], text summarization distills the essential information from a text concerning a task and user. The summary generated consists of 20% to 30% of the original content [4]. Extractive and Abstractive methods are two broad classifications of Text summarization. Abstractive summarization methods use natural language generation tools to

derive summaries from the original text. At the same time, extractive summarization methods use text mining approaches.

This paper presents the review of the recent literature on automatic text summarization developed for Indian languages. Text summarization for Indian languages is challenging due to the lack of adequate tools like appropriate taggers, parsers, synsets, etc. Therefore this paper focuses on the techniques used in the process of summarization. It also reviews various aspects like the type of documents (single/multiple), summary type (generic/query-based), summary characteristics (Extractive/Abstractive), etc. This literature reviews dataset used for experimentation and the evaluation metrics, etc. Finally, it briefly discusses the issues and challenges faced by researchers in the Indian language Text summarization field.

The order of the next sections of this paper is as follows. Section 2 describes the review on various text summarization techniques. Section 3 presents a review categorization of the text summarization process. The details of the data set are reviewed in section 4. Section 5 explains the evaluation metrics used by text summarization methods. Comparative study of Indian Language text summarization are explained in Section 6. Finally, the conclusion with issues and challenges, projected in section 7.

2. APPROACHES TO TEXT SUMMARIZATION

Depending on the earlier researchers' work, text summarization approaches are classified into four classes, namely statistical-based system, linguistic-based method, Machine learning-based methods, and hybrid systems [1], as shown in figure 1.

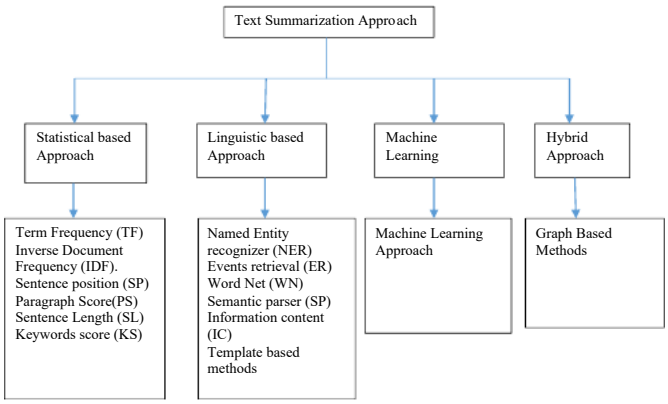

Figure 1 Classification of Text Summarization approaches used for Indian Languages

2.1 Statistical-Based Approaches

The statistical approaches score the sentence weights based on various parameters. The Term frequency shows how frequently a word has occurred in the document. If the sentence has the most commonly repeated words, those sentences are scored high [6]. Inverse document frequency gives rare word occurrences by calculating the log value of term frequency [7]. Sentence position is used to calculate the weight of the sentence and is referred to as position score. Position score is the position of a sentence in the document/ total number of sentences [9].

2.2 Linguistic Based Approaches

Linguistics of words incorporating lexical, semantic, and syntactic features of terms forms the keywords. Word dictionary, tree tagger, Parts of speech pattern, word n-grams are used for

lexical analysis of words [14]. Parsing builds the parse tree that helps to analyze the word syntactically [15].

In [16], linguistic parameters of information content are used to calculate the sentence score. This method is dependent on the corpus in which coherence among the sentences is obtained to generate the summary. In [17], template-based methods are used where the sentences' score is calculated based on the type of words like adjective, noun, verb, etc. In [18], maximum marginal reference finds the relevant sentence score concerning the word dictionary.

2.3 Machine Learning Approaches

Machine learning algorithms build the text summarization systems based on the training dataset. Various features are extracted to find the relevancy of the sentence in summary. Models such as Naive Bayes (NB)[20][21], Decision Trees (DT) [23], Hidden Markov model (HMM)[24], Neural Networks (NN)[19], support vector machines (SVM)[25][26], etc., are used to extract the relevant sentences. Features extraction plays a key role in building an efficient model.

2.4 Hybrid Approaches

These approaches combine the features of statistical, lexical, and machine learning based models. Graphical based methods and algebraic functions are used to identify the relevant score of a sentence. The top k highest appropriate scored sentences are selected to form a summary.

In [8], a graph-based approach is used to model the entire document. Each sentence in the text forms a graph's node and the weights calculated to the nodes. Sentence weight is the summation of the affinity weight of each word in the sentence. The affinity weight is calculated as the term frequency of a word/ total number of words in the document—the edges of nodes labeled with scores calculated using Levenshtein similarity weight between two sentences. Next, the vertex weight is calculated, which is the average of Levenshtein similarity weights of all edges connected to that vertex. The mean of sentence weight and vertex weight forms the sentence rank.

Depending on the classification shown in Figure 1, it is evident that the summarization methods fall into any one of these classes. Most of the Indian language text summarization methods developed keyword-based selection techniques. These techniques use statistical and linguistic-based approaches for the generation of Indian language summary. Machine learning techniques require data sets, an Indian language. But gold standard data sets for text summarization are not available for Indian languages. It raised the difficulty to measure the efficiency of the extracted summary.

3. CATEGORIZATION OF TEXT SUMMARIZATION PROCESS

Depending on the parameters used on which summaries are generated, the text summarization process is categorized into four groups, as shown in figure 2. The parameters used for the categorization are as follows.

- The number of documents from which the summaries are generated (Single/Multiple).
- Whether the summary generated is specific to a query or, in general.
- Characteristics of the summary generated involving extraction/abstraction.
- Summary obtained based on the type of learning methods adopted (Supervised/Unsupervised).

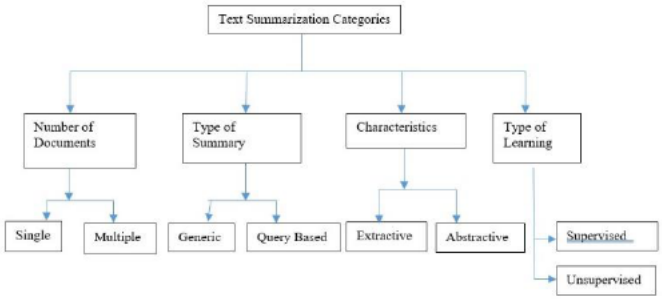

Figure 2 Categorization of Text Summarization process used for Indian Languages

3.1 Summary Depending on Input Documents

The number of input documents from which the summaries extracted is classified into a single document or multi-document summarization.

If a single document is given as input for summary extracted, such a method is referred to as single document text summarization [27] [28]. If more than one document is given as input to generate the summary, such a method is referred to as multi-document summarization [29] [30].

3.2 Characteristic Based Summary

Summaries are categorized into two types based on the characteristics by which the summary is obtained, namely extractive and abstractive.

Extractive Summarization summaries retrieve relevant sentences with the help of statistical and linguistic features in the input text [6][7][9][14]. These summaries select the sentences from the given document.

In contrast, abstractive summaries are generated by applying natural language understanding. Abstractive summaries depict the similarities of the way human beings usually write. The words in this summary may not present in the input text. Building an automation model for generating extractive summarization is relatively complex when compared to extractive methods [31].

3.3 Type of Summary

Summaries are categorized into two types based on the summary type, namely query based and generic. Keywords are identified by using statistical and linguistic methods. Sentences are ranked based on the number of keywords available in the text. Query focused summaries are generated by considering top-ranked sentences. The keywords should match with the words in the question [31]. Query-based summaries are applied to news articles to extract the outline of a specific news item.

In generic summarization, the summary is generated by identifying the sentences that have the most critical content. Sentences are ranked according to the lexical and semantic features [6][7][9]. Generic summaries are applied for the stories web pages and papers to understand the gist of the entire content [8]. Most of the Indian language text summarizations developed so far make are for generic summarization.

3.4 Learning Based Summary

Depending on the type of learning approach applied the summarization methods are classified into supervised and unsupervised.

In a supervised learning-based algorithm, the labeled data set is used to train the model to classify the relevant sentences. The statistical scores are used to mark the sentence as relevant or irrelevant—all the sentences labeled as relevant from the summary. The events and the events' temporal relations are identified in summary to arrange the sentences in the human-readable order [10].

In unsupervised learning based summaries, labeled data sets are not available. Instead, the clustering methods are applied to derive the summaries. The summaries that match a set of keywords are grouped to form the summary. In [33], the theme is determined using the K-means approach. A graph-based approach is applied for the selection of relevant sentences. Standard methods like page rank helped to generate the summary.

4. DATA SETS

In the English language, there are several standard data sets given by Data understanding conference (DUC), Text Analysis Conference (TAC), Computational Linguistics scientific document summarizer (CL-SciSumm), TISPER Text Summarization Evaluation Conference (SUMMAC)[34].

To test the language etc. But for Indian languages, there are no proper data sets available. Most of the researchers crawled the online news articles for experimentation of text summarization for Indian languages. In [16], data sets are collected from Indian languages from online news articles—70 news articles for the Hindi language, 50 articles for the Gujarati, 75 articles for Urdu.

In India there are 21 languages as per the records. Text summarization is attempted on very few languages like, Hindi, Bengali, Tamil, Punjabi, Guajarati, Marathi and Telugu [8][11][12][13]. All the methods developed have used their own data sets collected from online news for the experimentation purpose. Due to lack of sufficient lexical and syntactic rules. Very few researchers have attempted for text summarization for Indian Language documents. Building a dataset is a challenge for text summarization of Indian languages.

5. EVALUATION METRICS

Performance evaluation plays a significant role in evaluating the efficiency of the generated summary. Gold standard metrics for performance evaluation of English summaries are available. Data Understanding conference (DUC) has given the metrics like precision, recall, F measure, Similarity Score, and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores.

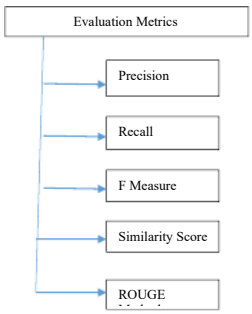

Figure 3 Evaluation Metrics used by Text Summarization of Indian Languages

The precision score is calculated based on the relevant sentences retrieved from the total number of sentences retrieved out of total sentences retrieved. The recall is defined as a relevant document extracted out of actually relevant. Depending on the precision and recall, F Measure

is calculated. Similarity scores are used to compare how relevant an automated summary is with respect to human-generated summaries. ROUGE score also determines summary quality by comparing with human summaries [3].

6. COMPARISON ON INDIAN LANGUAGE TEXT SUMMARIZATION TECHNIQUES

Table 1 Comparitive study on Indian Language Text Summarization

Language	AT				NoD		ST		SC		DT				EM	
	A1	A2	A3	A4	D1	D2	S1	S2	C1	C2	T1	T2	T3	T4	E1	E2
Urdu [5]	√				√		√		√		√				√	
Punjabi [36]	√	√			√		√		√		√				√	
Kannada [6][11][17][19]	√	√			√		√	√	√		√				√	
Bengali [7][13][33][35]	√		√		√	√	√		√		√	√		√	√	
Tamil [8][10]				√	√		√		√		√		√			√
Hindi [38][39]		√			√				√		√				√	
Malayalam [37]	√				√				√	√	√				√	
Odisa [12]	√				√				√		√				√	
Telugu [40]	√				√	√			√		√				√	

Table 2 List of Abbreviations used in Comparative Study of Indian Language Text Summarization

Approach Type (AT)	
A1	Statistical Approaches
A2	Linguistic Approaches
A3	Machine Learning Approach
A4	Hybrid Approaches
Number of Document (NoD)	
D1	Single
D2	Multiple
Summary Type (ST)	
S1	Generic
S2	Query based
Summary Characteristics (SC)	
C1	Extractive
C2	Abstractive
Database Type (DT)	
T1	Newspaper Articles
T2	Journal Articles
T3	Web pages
T4	Others
Evaluation Metrics (EM)	
E1	Precision, recall, F Measure
E2	ROUGE

Text summarization is attempted for some of the Indian languages like Hindi, Tamil, Punjabi, Odisa, Telugu, etc. The majority of the works have used statistical and linguistic-based approaches due to a lack of proper datasets, tools, other resources like stemmer, dictionaries, and synsets, etc. Most of the works are extraction where sentences are ranked and selected to obtain the summary. Abstractive methods for Indian languages is still in the starting stage. Annotated data sets need to be developed for Indian languages. The efficient summary extraction is dependent on the selection of appropriate features and their role in summary.

Standard evaluation metrics are used to compare the effectiveness and relevancy of the summaries generated. The comparative study table is provided in Table 1, which gives the overall picture of the various summarization techniques developed for different Indian Languages. (√) indicate the mapping of the parameter to the method. Table 2 gives the details of various text summarization parameters.

7. CONCLUSION

Text Summarization is a demanded application for users to obtain the gist of information in a short time of the search. Text Summarization for the English Language has started in Data Understanding Conference (DUC) since 2001. But in the Indian context, research in text summarization is slow due o the non-availability of appropriate tools and resources. This paper provides a survey on text summarization techniques developed for Indian languages. Various parameters like summary type, number of documents, approaches, etc., are used to compare different Indian languages. This survey gives an idea to bridge the research gaps in the research community involved in developing text summarization for Indian Languages. A few challenges and issues are highlighted for future work in this field. 1) Developing the resources like data sets, stop word lists, synsets for the Indian language like Telugu, Hindi, Tamil, etc. 2) Developing the multi-document summarization methods of Indian Languages. It included eliminating redundant sentences, building coherence, ordering in the summary sentences, etc. 3) Building a standard to identify the quality keywords that ensure better summary extraction.

REFERENCES

[1] C. Zhang, “Automatic keyword extraction from documents using conditional random fields,” Journal of Computational Information Systems, vol. 4 (3), 2008, pp. 1169-1180.

[2] E. Hovy, C.-Y. Lin, “Automated text summarization and the summarist system,” in: Proceedings of a workshop on held at Baltimore, ACL, 1998, pp. 197-214.

[3] I. Mani, M. T. Maybury, “Advances in automatic text summarization,” Vol. 293, MIT Press, 1999.

[4] G. Erkan, D. R. Radev, “Lexrank: graph-based lexical centrality as salience in text summarization,” Journal of Artificial Intelligence Research, 2004, pp. 457-479.

[5] Aqil Burney, Badar Sami, Nadeem Mahmood, Zain Abbas and Kashif Rizwan, “Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors” Pakistan International Journal of Computer Applications, 2012, (0975 – 8887) Volume 46–No.19.

[6] Jayashree.R. Srikanta Murthy. & Sunny.K. "Document summarization in kannada using keyword extraction", 2011 CS & IT-CSCP, pp. 121-127.

[7] K. Sarkar, “An approach to summarizing Bengali news documents,” In proceedings of the International Conference on Advances in Computing, Communications and Informatics, 2012 ACM, pp. 857-862.

[8] Sankar K, Vijay Sundar Ram R and Sobha Lalitha Devi, Text Extraction for an Agglutinative Language, Problems of Parsing in Indian Languages, M a y 2 0 1 1 Special Volume, PP 56-59.

[9] Krish Perumal, Bidyut Baran Chaudhuri, Language Independent Sentence Extraction Based Text Summarization, Proceedings of ICON-2011: 9th International Conference on Natural Language Processing, pp.213-217.

[10] Jagadish S KALLIMANI, Srinivasa K, G, Information Retrieval by Text Summarization for an Indian Regional Language, 2010 IEEE. 10.1109/NLPKE.2010.5587764.

[11] Jayashree.R, Srikanta Murthy.K and Sunny.K, Document summarization in kannada using keyword extraction, International Journal of soft computing vol 2 no 4 , Nov 2011, pp 81-93.

[12] R. C. Balabantaray, B. Sahoo, D. K. Sahoo, M. Swain , Odia Text Summarization using Stemmer, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Volume 1– No.3, February 2012

[13] Kamal Sarkar Bengali text summarization by sentence extraction, Proceedings of International Information Management(ICBIM-2012),NIT Conference on Business and Durgapur, PP 233-245.

[14] R. Barzilay, M. Elhadad, "Using lexical chains for text summarization," Advances in automatic text summarization, 1999, pp. 111-121.

[15] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in: Proceedings of the 2003 conference on Empirical methods in natural language processing, ACL, 2003, pp. 216-223.

[16] Alkesh Patel, Tanveer Siddiqui, U. S. Tiwary , A language independent approach to multilingual text summarization, RIAO2007, Pittsburgh PA, USA, May 30- June 1(2007), pp,123-132.

[17] Embar, V.R., Deshpande, S.R., &Vaishnavi, A.K.. "sArAmsha- A Kannada Text Summarizer", IEEE., Advances in computing, ICACCI, International Conferencence on 22-25 Aug 2013. Pp.540-544.

[18] Dhanya, P. M., and M. Jathavedan, (2013). NCILC seminar proceedings.

[19] Jayadhree R, Srikantamurtyhy K, Basavaraj,S Anami, Vijay M, and Bharathi B N, " An Artificial Neural network approach to text summarization for the south Indian languages of kannada",IEEE, Hybrid Intelligence System (HIS), 13th International Conference, 4-6 2013,pp 45-48.

[20] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," In Proceedings of tenth European Conference on Machine Learning, 1998, Springer- Verlag, pp. 4–15.

[21] J. D. M. Rennie, L. Shih, J. Teevan, D. R. Karger, "Tackling the poor assumptions of nave Bayes classifiers," In Proceedings of International Conference on Machine Learning, 2003, Pp. 616–623 .

[22] H. L. Chieu, H. T. Ng, "A maximum entropy approach information extraction from semi-structured and free text," In Proceedings of the Eighteenth National Conference on Artificial Intelligence, American Association for Artificial Intelligence, 2002, pp. 786–791.

[23] L. Rabiner, B. Juang, "An introduction to hidden Markov model," Acoustics Speech and Signal Processing Magazine, vol. 3(1), 2003, pp. 4–16.

[24] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines" mach. learn. Machine Learning, vol. 46(1–3), 2002, pp. 389–422.

[25] L. N. Minh, A. Shimazu, H. P. Xuan, B. H. Tu, S. Horiguchi, "Sentence extraction with support vector machine ensemble," In Proceedings of the First World Congress of the International Federation for Systems Research, 2005, pp. 14-17.

[26] D. Marcu, Discourse trees are good indicators of importance in text, Advances in automatic text summarization (1999) 123-136.

[27] S. M. Harabagiu, F. Lacatusu, Generating single and multi-document summaries with gistexter, in: Document Understanding Conferences, 2002, pp. 40-45.

[28] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, C. A. Mehdiyev, Mcmr: Maximum coverage and minimum redundant text summarization model, Expert Systems with Applications 38 (12) (2011) 14514-14522.

[29] Z. L. Min, Y. K. Chew, L. Tan, "Exploiting category-specific information for multi-document summarization," in Proceedings of COLING, ACL, 2012, pp. 2093–2108.

[30] P.-E. Genest, G. Lapalme, "Framework for abstractive summarization using text-to-text generation," in: Proceedings of the Workshop on Monolingual Text-To-Text Generation, ACL, 2011, pp. 64-73.

[31] S. Fisher, A. Dunlop, B. Roark, Y. Chen, J. Burmeister, "Ohsu summarization and entity linking systems," in: Proceedings of the text analysis conference (TAC), Citeseer, 2009.

[32] A. Das and S. Bandyopadhyay, "Topic-Based Bengali Opinion Summarization", International Conference COILING '10, Beijing, pp. 232–240, 2010.scisumm-corpus @ Https://Github.Com/Wing-Nus/Scisumm-Corpus.

[33] N. Uddin and S. A. Khan, "A Study on Text Summarization Techniques and Implement Few of Them for Bangla Language", Proceedings of international Conference on Computer and information technology, 2007, IEEE, pp. 1-4.

[34] Vishal Gupta, Gurpreet Singh Lehal, "Features Selection and Weight learning for Punjabi Text Summarization", International Journal of Engineering Trends and Technology-2011 Volume2 Issue2.

[35] Ajmal E.B, Posna P Haron, (2015) "Summarization of Malayalam Document Using Relevance of Sentences" International Journal of Latest Research in Engineering and Technology, Volume I Issue 6 pp 08-13.

[36] Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T. (1983). Statistical semantics: Analysis of the potential performance of key-word information systems. Bell System Technical Journal, 62(6), 1753-1806.

[37] U. Garain, A. K. Datta, U Bhattacharya and S.K. Parui, (2006), Summarization of JBIG2 Compressed Indian Textual Images,‖ Proceeding of 18th International Conference onPattern Recognition (ICPR,,06), IEEE, Kolkata, India, Vol. 3, Pp. 344-347, 2006.

[38] Reddy Naidu, Santosh Kumar Bharti, Korra Sathya Babu and Ramesh Kumar Mohapatra, Text Summarization with Automatic Keyword Extraction in Telugu e-Newspapers, Springer Smart Computing and Informatics, Smart Innovation, Systems and Technologies 77, 2018, pp: HTTPS://DOI.ORG/10.1007/978-981-10-5544-7_54. PP 543-555

BIOGRAPHY

First Author Mr.Kishore Kumar Mamidala is Associate Professor in the Department of Computer Science at Vivekananda Institute of Technology & Science, Karimnagar since June 2008. Prior to this he was an Assistant Professor in the CSE department at Dr.VRK College of Engineering since 2006. He received B.Tech in CSE from JPNCE (affiliated to JNTUH), M.Tech. in Computer Science from JBIET (affiliated to JNTUH) in 2008, and pursuing Ph.D. in Computer Science from the JNTU, Hyderabad.

Mr Kishore Kumar has been involving in research activities to develop the systems which can detect useful, new, and timely sentence-length updates about a developing event. In Temporal Summarization, he works on extracting key sentences to produce the automatic text summaries of the input documents, and hence is a potential solution to the information overload problem. He has published 10 research articles and also a book chapter on his research activities titled " Mining Behavioural Data: Data Mining Technologies using Machine Learning Algorithms".

Second Author Dr. Suresh Kumar Sanampudi is presently working as Assistant Professor and Head in the Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad. His research interests include Artificial Intelligence, Natural Language Processing, Information Retrieval, Algorithms and Information Security.