National Conference on Recent Advances in Computer Science and IT (NCRACIT)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 1 | ISSN: 2456-3307

A Comprehensive Study of Text Summarization Algorithms

Yash Dhankhar[1], Indu Bala[2], Swati Singh[3], Sunil Dalal[4]

[1]Computer Science and Engineering, BabaMastnath Engineering College, Rohtak, Haryana, India
[2]Information Systems, Delhi Technological University, Delhi, India
[3]Computer Science and Engineering, Delhi Technological University, Delhi, India
[4]Assistant Professor, BGSB University, Rajouri, J&K,India
yashdhankhar92@gmail.com[1], induyadav330@gmail.com[2], swatisngh337@gmail.com[3], sunildalal57@gmail.com[4]

ABSTRACT

This document provides some minimal guidelines (and requirements) for writing a research paper. Issues related to the contents, originality, contributions, organization, bibliographic information, and writing style are briefly covered. Evaluation criteria and due dates for the research paper are also provided.

Keywords : Research Paper, Technical Writing, Science and Technology

I. INTRODUCTION

Text summarization is a method to produce a concise and important piece of information from a larger set of text which can be a text document, an article or a blog. Text Summarization aims to provide a summary of given text while preserving its information and intent. The summary is a small piece of information that describes a set of paragraphs or documents. Summary generated is generally less than forty percent of the original text data and it should be even less than that in the case of large datasets. The summary should retain the important data present in the document, should be controllable, short and succinct. Summarization of text data is done in many ways depending upon the various parameters based on the position and format of words and sentences. Automatic Text Summarization [1] accumulates the data from several documents to present the final shorter piece of information as a result, which is shorter, informative and preserves the real intent of information. These small summarized versions save valuable time by presenting unambiguous important information. With the increasing amount of digital data, it has become difficult to retrieve the needed and concise information. Automatic text summarization caters to the very need of the time. There are methods, which are helpful to produce a summary. First Division, which categorizes the summarization approaches, is based on the content of the summary produced. There are two approaches-Extraction and Abstraction [2]. As the name suggests, Extraction is domain independent, it mainly aims at finding out the important sentences and later presenting a set of important sentences as Summary. On the contrary, Abstraction is domain dependent, it processes the available information and new sentences are prepared by understanding the content, also considers human knowledge by preparing the goal to produce a summary.

Applications of Text Summarization: Automatic text summarization has many important applications. One of the important application is in Medical area, where a lot of unclassified information is available and many times a medical associate is required to find about some information specific to a medical condition for research or diagnostic purpose from the large heap of documents. Finding out relevant information involves the reading of numerous documents and problem/patient's records. Summarization specifically personalized to the medical area is very useful, as it not only saves time but increases the efficiency of a medical expert.In legal processes, a typical case study involves consideration of loads of information consisting of law expertise books and numerous related previous judicial case studies, thus leading to an overload of information. The legal experts perform a tedious and responsible task, their time and resources are expensive. To find out an important piece of information unambiguously and in less time is desirable to cater the needs of fast and correct court decisions. For Research Purposes, hundreds of research papers need to be considered in any research domain to find out a specification. In this way to know what lies inside of the paper, researchers need to read more than the abstract but less than the paper, so summarization may be applied to get the customized summary of the content by applying the desirable method. On the internet, Summarization is used in multiple applications. Various newspaper sites and related apps provide everyday news with the use of summarization in order to save time and space while keeping the important key information. Mainly editorials are summarized while keeping the intent of author intact. Further, there are also applications for mobile devices like smart phones, tablets etc. they include small screen area and time available to read. In the corporate world, 'meeting minutes' need to be read in small time and associated documents need to be looked into before next meeting without the help of human and other resources. For blind people, a lot of time can be saved by readers while reading to them by giving them an important piece of information instead of the whole document. The paper is structured as follow: Section 2 gives an insight into the related work done in text summarization. Section 3 gives an overview of automatic text summarization. Section 4 concludes the paper and gives the future scope in this research area.

II. RELATED WORK

A lot of good work has been done on the text Summarization. In most of the Summarizers, Sentences are considered as a feature vector[] and various algorithms are applied depending upon the position of Sentences, Vocabulary intersections, title distribution and the type of data. Apart from the sentence related data other features include the structure of the document and popularity of the content. Most of the work has been done in the extraction domain, but various different ideas have been explored like multiple document summaries, language based summarizers etc. In 1955, Henry Peter Luhn, IBM inventor first published a paper entitled 'A new method of recording and searching information' (Luhn, 1953) [3]. He developed many Information retrieval applications. Later in 1969, Edmundson described a new extraction method based on extraction using three components: pragmatic words(cue words), structural indicators and topic heading words[4]. In 1980's AI methods came into existence for summarization using hybrid approaches for different types of summarization i.e. multiple documents, multimedia etc. One of the applications is KWIC (Keyword in Context) by using three fundamental elements: Keyword, title, and context. In the last two decades various new and hybrid methods have been described. TextRank[5], cluster based[6], Rhetoric based[7], Topic models[8], ILP based method[11] etc.

In an Abstractive domain, as new sentences need to be designed, it needs a deeper analysis of the original text information. It involves an understanding of the text by linguistic methods[9] to provide an interpretation to match the level of human generated summary. There are two main approaches for doing this i.e. structure based and semantic based approach. In Structure based approach, most weighted data is encoded by cognitive schemas[10]. Structures such as a tree, ontology, lead and body phrase structures are the schemas mostly used for structured approach. In the second approach, semantic based uses a Natural Language generation system to process the semantic information to categorize the grammar variants such as noun and verbs by processing linguistic data. To achieve a true abstractive summarization is still a dream.

III. OVERVIEW OF AUTOMATIC TEXT SUMMARIZATION

A. Features

In order to decide the degree of importance a sentence to include or exclude from the final summary formation, a list of features used by the researchers are listed below:

1) Term Frequency [19]: Frequency of a word is measured for the whole source document. Then, the scores are assigned to each sentence based upon the number of frequent words belonging to the particular sentence. Sentences with highest weights are considered for final summary. TF IDF is widely used for calculation of word frequency.
2) Location: In a text document, the position of a sentence also tells about its relevance in the summary. While calculating the weighted Score of a sentence certain sentences are weighted higher than others.
3) Cue Method: Sentences including words that adds to limitations or advantages of the content are weighted high i.e. “in summary”, “significantly”, “describes” “concludes” are cue words.
4) Title/ Headline[20]: words included in the topic or theme of the content are considered relevant. Sentences which includes these words are assumed to be important for the summary. Some constant weight is added while weight calculation.
5) Sentence Length[19]: Number of words in a sentence defines the length of sentence, which is a factor in deciding sentence relevance in the final summary.
6) Proper noun [19]: Source document sentences, which include proper nouns, are assumed suitable for the final summary.
7) Proximity [19]: To identify relations among words or entities, distance between them is considered an important factor.
8) Similarity: To find out the relevance of a sentence in a document we calculate similarity among sentence and other sentences of the source document.

B. Summarization Approaches

The summarization can be performed on single documents and multiple documents as well.

- Single Document Summarization: In single document summarization[16], one source document is analyzed and processed to generate a quality summary. Both the approaches, extractive as well as abstractive can be applied on a single text document.
- Multi Document Summarization: Multi document summarization is a technique which involves the information extraction from more than one document. Multiple source documents are analyzed and evaluated to generate an important and non-redundant piece of information. Multiple document summarization [17] technique came after single document summarization to cater the needs when we need to concise data, which is distributed in multiple files. News on the internet is based on the web based clustering systems.

C. Methods

Two methods, extractive and abstractive are discussed here.

1) Extractive Summaries

Extractive summaries are simple to form as they only include few important sentences from the text document. They decide the importance of sentences in a document and decides to include the most informative sentences, paragraphs etc. in the final summarized result.

2) Abstractive Summaries

Abstractive summaries are prepared with a combination of newly formed sentences by analyzing a set of important information. New formed sentence should be coherent and complete. Abstractive summaries are generated by proper understanding the source document and then forming new sentences. It produces a representation of internal semantic details, then uses the natural language techniques for the final summary generation.

D. Types

Query based summaries and generic summaries are the two types of summarization types.

1) Query Based Summaries

In Query based summaries [18]; the final summary is generated based on query raised by a user. This technique can be applied on the single document as well as multiple documents. The relevance of a sentence for the final summary result is calculated based upon the frequency of words in a document. A sentence in the original document, which includes the keywords provided in a query by the user, is scored high than others. Sentences with the high scores are suitable for final summary.

2) Generic Summaries

Generic summaries provide a complete review of the source document unlike query based technique only caters to the query of the user. For the content overview, generic summaries are suitable. This aims to identify the key topics and decrease the redundancy to a possible minimum. Generic summaries categorize and describe the main idea of the source content.

E. Summary Techniques

Four well-known techniques are used in summarization.

- Semantic and Syntactic (Rule-based)
- Statistical Technique
- Clustering Technique
- Machine Learning Technique

1) Semantic and Syntactic Analysis

Rule-based technique is used to find and present the association among different sentences by applying on source content for text summarization. These can be categorized as following:

- Graph Representation
- Lexical Chains
- NLP (Natural Language Processing)

- The graph representation is done during summarization by lexical graphs, sentences are represented as Weighted graphs, unweighted graphs, graph matching etc are tasks performed during summarization process.
- Lexical chains are used for building chains of identified units for summarization with the help of co-reference chains and lexical semantics etc.
- Natural Language information processes language data to extract information also uses part of speech for summary production. There are two techniques for summarization under Natural Language Processing listed below:

1) Plain text Summarization
2) Multilingual Summarization

In Plain text summaries, resultant summary is in the same natural language but in multilingual text summarization [21] resultant summary may be in one of the languages in which sources are written or may be a totally different language. It came into existence in 2005. Evans (2005) described the need of summarization in a particular language from different sources available in different languages.

This technique is still in research stage but it has features that are very useful in news reporting where data is combined from different foreign news agencies and summarized in language used in a region.

2) Statistical Technique

For extraction of relevant information, some systems use Statistical Techniques. This technique uses statistical methods generally applied with Binomial Distribution, sentence compression and calculated scores. This technique is used by Hidden Markov model.

Conroy and O'Leary [24] employed statistical technique by hidden Markov model approach for summarization of plain text documents. A sequential Model was prepared for the evaluation of local independence.

This system has three key parameters as the length of the sentence in processing, the position of the sentence in document and likeliness of key terms in the sentence being evaluated.

3) Clustering Technique

When multiple objects are grouped together based upon their properties and characteristics, this process is termed as Clustering[6]. A cluster consists of the objects having similar properties. In text summarization, we use clustering to group similar type of sentences together. In a document different topics are arranged in a specific ordering. In this technique, firstly clusters are generated and then sentences are selected. The sentence is also chosen based upon the location or position of the sentence in a document. A score of a sentence increases if it has multiple occurrence hence higher probability of selection in the final summary.

4) Machine Learning Technique

Automatic text summarization can be effectively done by Machine learning techniques. Some of the machine learning approaches are discussed as follows.

a) Naive Bayes Approach : Kupiec (25) described a method for summarization in which a classification function known as naïve Bayes classifier is used which is responsible for the each sentence to be a part of the summary.
b) Rich features and Decision Trees: Lin and Hovy (27) used "sentence position" in which a weight is provided to sentence based upon its position in the text. This method is also known as position method.
c) Log Linear Models : Osbrone (2002) described the Log Linear model approach[28] for the plain text summarization. This approach is different than the previous approaches which always assumed feature independence. The system showed that this approach is better than naïve Bayes classifier approach.
d) Neural Networks : Svore (29) produced an algorithm based upon neural networks and used the third party features like dataset to resolve the problem of extractive summarization.

F. Comparative Study

As Text summarization first approach came in the 1950s since then many new approaches and techniques have been implemented and exercised. Different techniques involve a specific set of feature selection and the content on which the algorithm is applied, We have done a comparative analysis of few techniques in the table listed below:

Author/Year	Method	Features/ Content Selection	Technique Used	Summarization Approach
1995 Julian Kupiec[25]	algebraic method	like length, the position of words, uppercase words	using a naïve-bayes classifier	Extractive Summarization
1997 ChinYew Lin[41]	algebraic method	the position of sentences	Rich Features and decision trees	Extractive Summarization
1999 Eduard Hovy[31]	symbolic word knowledge	concepts relevancy	NLP processing	Single Document Summarization(A)
2005 S.P Yong[42]	Text pre-processing and	Keywords Extraction	used neural networks	Abstractive Summariz

	subsystem	Summary production		ation
1984 Ruqaiya Hasan[43]	Coherence relation	similarity chains	lexical cohesion	Single Document Summarization(A)
1988 William C.Mann [44]	Tree based	to encode the terminal nodes of a tree	RST (rhetorical structure theory)	Abstractive Summarization
1997 Branimir Boguraev[45]	Saliency based content characterization	rank the important sentences	Ranking algorithm	Extractive Summarization
2010 Li Chengcheng [46]	rhetoric relations	candidate sentence	RST (rhetorical structure theory)	Abstractive Summarization
Xiaojun Wan in 2008 [47] used graph based method by introducing	used graph based method	The two-link graph for both sentences and documents	Graph based method	Multiple Document Summarization(A)
2012 Tiedan Zhu [48]	Sentence closeness Parameter	Logical closeness to document	Sentence Co-relation Method	Multiple Document Summarization€

IV. CONCLUSION

Automatic Text Summarization is used to get an important piece of text from a larger document. A large number of algorithms designed and implemented to get a good, coherent and non-redundant summary a little similar to the human prepared summary. Simple single document extractive algorithms have given better results in different domains as compared to abstractive summarization algorithms. Extractive summarizers are used to select the important set of sentences from the source document based on top scoring Sentence-ranking method. Although, By performing Automatic Text Summarization to get a gist of the input text documents equivalent to human interpreted summary is not yet fulfilled, but by improving the existing algorithms, the value of evaluation metrics is increasing. With the rapid increase in the electronic data on internet and less time to read the documents based on a similar topic has called a need to design accurate and efficient Multi- document summarization systems. As research on text summarization started 50 years ago and a lot of work has been done in the extractive area in both the single and multiple document domains but there is still a long path to cover in this field. Abstractive summarizers aim to import more information in a single sentence rather than include the sentence as a whole. Multi-document Abstractive Summarization is the area which is needed to be explored.

Over time, attention has drifted from summarizing scientific articles to news articles, electronic mail messages, advertisements, and blogs. Domain associated summarizes can be a solution to get more accurate summaries. Medical and Legal matters domain can be highly benefitted from this area of research even if they focus only on small details related to a general summarization process and not on building an entire domain dependent summarization system.

II. REFERENCES

[1] D.K. Gaikwad and C.N. Mahender "A Review Paper on Text Summarization", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016,154-160

[2] Radev, D. R., Hovy, E., and McKeown, K. (2002) "Introduction to the special issue on summarization." Computational Linguistics., 28(4):399-408

[3] Luhn, H. P. (1958) " The automatic creation of literature abstracts". IBM Journal of Research Development, 2(2):159–165

[4] Edmundson, H. P. (1969) " New methods in automatic extracting". Journal of the ACM, 16(2):264–285.

[5] R.Mihalcea, and P.Tarau, "TextRank: Bringing Order into Texts." In Proceedingsof Empirical Methods in Natural Language Processing (EMNLP). pp. 404-411. 2004.

[6] Z.Pei-ying, and L.Cun-he, "Automatic Text Summarization based on Sentences Clustering and Extraction," Proceeding of the 2nd IEEE International Conference on Computer Science and Information Technology. pp. 167-170. 2009

[7] 20IOy International Conference on Computer Application and System Modeling (ICCASM 2010) Automatic Text Summarization Based On Rhetorical Structure Theory Li Chengcheng 595-598[8] D. Blei, A. Ng, and M. Jordan " Latent Dirichlet allocation". In Journal of Machine Learning Research, 3:993–1022, January2003.

[8] Barzilay, R. and Elhadad, M. (1997). "Using lexical chains for text summarization." in Proceedings ISTS'97. pg. 38-41

[9] Radev, D. R. and McKeown, K. (1998) "Generating natural language summaries from multiple on-line sources." Computational Linguistics, 24(3):469–500

[10] S. Banerjee, P.Mitra and K. Sugiyama " Multi-Document Abstractive Summarization Using ILP Based Multi-Sentence Compression"in Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

[11] Lin, C.-Y. (2004). "Rouge: A package for automatic evaluation of summaries." In Proceedings of the ACL-04 Workshop, pages 74–81, Barcelona, Spain

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu "BLEU: a Method for Automatic Evaluation of Machine Translation" in Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318

[13] S. Brin and L. Page "The PageRank Citation Ranking:Bringing Order to the Web" in 1999

[14] Mc Keown, K. R. and Radev, D. R. (1995). "Generating summaries of multiple news articles." in Proceedings of SIGIR '95, pages 74–82, Seattle, Washington.

[15] Jagadeesh J, Prasad Pingali, Vasudeva Varma "Sentence Extraction Based Single Document Summarization" Workshop on Document Summarization, 19th and 20th March, 2005, IIIT Allahabad

[16] Kamal Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", TECHNIA – International Journal of Computing Science and Communication Technologies, vol. 2, no. 1, Jul. 2009.

[17] F. Canan Pembe and Tunga Güngör, "Automated Query-biased and Structure-preserving Text Summarization on Web Documents," in Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, İstanbul, June 2007.

[18] Reeve Lawrence H., Han Hyoil, Nagori Saya V., Yang Jonathan C., Schwimmer Tamara A., Brooks Ari D., "Concept Frequency Distribution in Biomedical Text Summarization", ACM 15th Conference on Information and Knowledge Management (CIKM), Arlington, VA, USA,2006.

[19] Khan Atif, Salim Naomie, "A review on abstractive summarization Methods", Journal of Theoretical and Applied Information Technology, 2014, Vol. 59

[20] Evans, D. K. (2005). "Similarity-based multilingual multi-document summarization." Technical Report CUCS-014-05, Columbia University.

[21] Edmundson, H. P. (1969). "New methods in automatic extracting." Journal of the ACM, 16(2):264–285.

[22] Martins, Camilla Brandel and Lucia Helena Machado Rino. "Revisiting UNLSumm: Improvement Through a Case Study." (2002).

[23] Conroy, J. M. and O'leary, D. P. (2001). "Text summarization via hidden markov models." In Proceedings of SIGIR '01, pages 406–407, New York, NY, USA

[24] Kupiec, J., Pedersen, J., and Chen, F. (1995). "A trainable document summarizer." In Proceedings SIGIR '95, pages 68–73, New York, NY, USA.

[25] Aone, C., Okurowski, M. E., Gorlinsky, J., and Larsen, B. (1999). "A trainable summarizer with knowledge acquired from robust nlp techniques".pages 71–80

[26] Lin, C.-Y. and Hovy, E. (1997). "Identifying topics by position." In Proceedings of the Fifth conference on Applied natural language processing, pages 283–290, San Francisco, CA, USA.

[27] Osborne, M. (2002). Using maximum entropy for sentence extraction. In Proceedings of the ACL'02 Workshop on Automatic Summarization, pages 1–8, Morristown, NJ, USA

[28] Svore, K., Vanderwende, L., and Burges, C. (2007). "Enhancing single-document summarization by

combining RankNet and third-party sources." In Proceedings of the EMNLP-CoNLL, pages 448–457.

[29] Barzilay, R. and Elhadad, M. (1997). "Using lexical chains for text summarization." in Proceedings ISTS'97.

[30] Hovy, E. and Lin, C. Y. (1999). "Automated text summarization in summarist." In Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, pages 81–94. MIT Press

[31] N. Aletras and M. Stevenson. "Evaluating topic coherence using distributional semantics." In Proc. Of the 10th Int. Conf. on Computational Semantics (IWCS'13), pages 13–22, 2013.

[32] Kamal Sarkar "Automatic Single Document Text Summarization Using Key Concepts in Documents" J Inf Process Syst, Vol.9, No.4, pp.602-620, December 2013

[33] I. Chen "Integer Linear Programming Models for Constrained Clustering" in International Conference on Discovery Science 2010: Discovery Science pp 159-173

[34] Günes Erkan and Dragomir R. Radev. 2004. "LexRank: graph-based lexical centrality as salience in text summarization". J. Artif. Int. Res. 22, 1 (December 2004), 457-479.

[35] Jinqiang Bian, Zengru Jiang, Qian Chen 2014 "Research On Multi-document Summarization Based On LDA Topic Model" Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics 113-116

[36] Virendra Kumar Gupta Tanveer J. Siddiqui "Multi-Document Summarization Using Sentence Clustering" IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction, Kharagpur, India, December 27-29, 2012

[37] The Porter Stemming Algorithm [Online] Available:http://tartarus.org/~martin/PorterStemmer

[38] George A. Miller. "WordNet: A Lexical Database for English." Communications of the ACM, pages 39-41, November 1995

[39] Sherry and Dr. P. Bhatia " A Survey to Automatic Text Summarization Techniques" International Journal of Engineering Reasearch, October 2015 Pg. 1045- 1053

[40] Chin-Yew Lin and Eduard Hovy, "Identifying Topics by Position," In Proceedings of the Fifth conference on Applied natural language processing, San Francisco, pp. 283-290, 1997.

[41] S. P. Yong, A. I. Z. Abidin and Y. Y. Chen, "A Neural Based Text Summarization System," 6th International Conference of Data Mining, pp. 45-50, 2005.

[42] Ruqaiya Hasan, Coherence and Cohesive Harmony, In: Flood James (Ed.), Understanding Reading Comprehension: Cognition, Language and the Structure of Prose. Newark, Delaware: International Reading Association, pp. 181-219, 1984.

[43] William C. Mann and Sandra A. Thompson, Relational Propositions in Discourse, Defense Technical Information Center,

[44] Branimir Boguraev and Christopher Kennedy, "Saliencebased Content Characterization of Text Documents," In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, 1997.

[45] Li Chengcheng, "Automatic Text Summarization Based On Rhetorical Structure Theory," International Conference on Computer Application and System Modeling (ICCASM), vol. 13, pp. 595-598, October 2010.

[46] Xiaojun Wan, "An Exploration of Document Impact on Graph-Based Multi-Document Summarization," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,

[47] Tiedan Zhu and Xinxin Zhao, "An Improved Approach to Sentence Ordering For Multi-document Summarization," IACSIT Hong Kong Conferences, IACSIT Press, Singapore, vol. 25, pp. 29-33, 2012.