Journal of Engineering and Applied Sciences 12 (23): 7194-7200, 2017
ISSN: 1816-949X

A Study on Summarization of Technical Documents

[1]Syed Sabir Mohamed and [2]Shanmugasundaram Hariharan
[1]Faculty of Computer Science and Engineering, Sathyabama University,
Chennai, Tamil Nadu, India
[2]Department of Information Technology, Vel Tech Multi Tech, Chennai, India

Abstract: The popularity of internet and availability of online resources has increased the demand for generating precise summary to be extracted from single or multiple sources. This has also necessitated the need for intensive research in the area of automatic text summarization. Over several decades, the knowledge scarcity problem has been addressed in different perspectives with varied domains and using different paradigms. This study intends to investigate some of the most relevant approaches for summarizing technical documents emphasizing empirical methods and extractive techniques for generating summaries of technical documents. The results were also tested with commercial summarizers and the investigation seems to be promising in terms of precision and recall.

Key words: Summarization, stop words, Gist, technical document, extraction, sentence scoring

INTRODUCTION

Summarization denotes proving statements according to user specified target ration, i.e., main points of some information. Today's modern era has provides voluminous information to be shared online. But there are several shortfalls like repetition of data. It is always better to have summary of something rather than a long description about something. Summarization is therefore, has potential significance in recent years (Mohd *et al*., 2016). Earliest instances of research on summarizing scientific documents proposed paradigms for extracting salient sentences from text using features like word and phrase frequency (Luhn, 1958), position in the text (Baxendale, 1958) and key phrases (Edmundson, 1969). Popular approaches for text summary generation are statistical approach, text-connectivity scheme, graph-based scheme, algebraic and non-extractive approaches (Patil *et al*., 2013).

The task of technical document summarization is to take an information source, extract the content from it and present the most vital content to the user in a condensed form. It must also be in accordance to the user or application's needs. This area is highly interdisciplinary and relates to natural language processing, artificial intelligence information retrieval and information extraction (Mallett *et al*., 2004). The main difference between automatic and human-based text summarization is that humans can capture and convey subtle themes that permeate documents whereas automatic approaches have a large difficulty to do the same. Nonetheless as the amount of information available in electronic format continues to grow, research into automatic text summarization has taken on renewed interest with investigation on different parameters (Mohamed and Hariharan, 2016). Several summary generation methods have been investigated in the recent past. The methods presented highlight the comparative points between those techniques. Also, different types of summaries generated by using different approaches and methods have been investigated (Tabassum and Oliveira, 2015). Some commercially available extractive summarizers like Copernic* and Word+ use certain statistical algorithms to create a list of important concepts and generates summary accordingly.

The ability to summarize information automatically and present results to the end user in a compressed, yet, complete form would save much time. A good summarizer would be a great boon to the student community who can capture the central idea of an scientific study without the need to read it thoroughly. The main objective of the proposed work is to present a brief overview of the technical paper in the form of minimal summary (one or two page) and standard summary. It is in general if a summary includes pictures, tables in the summaries, then understanding of the study would be very easy. We have presented a summarizer framework called as "TechSumm" which could present a precise summary from technical document content based on the user choice.

Considering the rapid growth of volume of scientific literature documents, several technique enables for scientific research articles with the extension of scientific

Corresponding Author: Syed Sabir Mohamed, Faculty of Computer Science and Engineering, Sathyabama University, Chennai, Tamil Nadu, India

knowledge and practical usages. Although, there have been several efforts to extract informative summaries and patent from research articles, there are few attempts in other scientific literatures. This study proposes a framework generation of summaries from scientific documents.

Literature review: Researchers and scientists increasingly find themselves in position to quickly understand and explore methods for implementing technical materials for research. Due to the vast amount of scientific literature in each field of study, people face lot of troubles in reading though the scientific literature documents as a whole. One possible approach is to produce a summary of the scientific topic. Few models have been proposed for summarizing the scientific articles which can be a part or the entire document an entire topic. The model is based on analyzing others viewpoint of the target articles contributions and the study of its citation summary network using a clustering approach (Qazvinian and Radev, 2008). The proposed approaches also need to outperform several baselines in terms of both extraction quality and fluency (Jbara and Radev, 2011). Factors like diversity, readability, cohesion and ordering of the sentences included in the summary need to be thoroughly considered which would otherwise lead to summaries which would be noisy and confusing.

Text summarization for different domains is more successful like Estonian news paper text summarizer named as Estum (Muurisep and Mutso, 2005). Such summarizers adopt procedures in assigning weighs for sentences in documents based on position, format and keywords. The summarizers should also be able to deal with a variety of document formats such as HTML and XML. They must also be able to exploit information in the tags associated with these documents (Jing, 2000). Summary generation are categorized as “Indicative summary” which gives an idea of what the text is about without conveying specific content and “Informative summary” which provides some shortened version of the content (Hahn and Mani, 2000).

Summarizing scientific texts is a language independent and generic extractive process of extracting qualitative summary from scientific documents. GistSumm (Filho and Pedro, 2007) is developed for producing Gist from text documents which are highly technical in nature. The system is an easy to use user friendly and has been tested for its effectiveness.

Summarization of scientific articles concentrating on the rhetorical status of statements provides summaries that can highlight the new contribution of the source article. The summaries generated of such kind involve computational linguistics analysis and human judgments. The algorithm involves our measuring judge’s agreement of annotations, training and classification from fixed set of rhetorical categories yielding a single-document summary which is task-oriented and user-tailored (Teufel and Moens, 2002).

With the automatic semantic annotation framework, approach including tag set modeling for semantic annotation, semi-automatic annotation tool, manual annotation for training data preparation and supervised machine learning were developed (Jung, 2017) *http://www.copernic.com/en/products/summarizer +https://www.groovypost.com/howto/summarize-articles-mi crosoft-word/.

The design and experiments involve two different domains such as information and communication technology and chemical engineering. In addition, three application scenarios of annotation framework were used which serves as a guide for potential researchers who are willing to link their own contents with external data.

Extractive text mining in collaboration with summarization techniques were found to be useful for generating precise summaries for scientific literatures. In order to generate summaries, citations were used (Qazvinian *et al*., 2013) summaries. The researchers have proposed C-LexRank, a model for summarizing single scientific articles based on citations which employed community detection and extracts salient information-rich sentences. The experimental investigation employed set of papers which covers the same scientific topic. Based on this, extractive summaries based on the set of Question Answering (QA) and Dependency Parsing (DP) was obtained.

Evaluation of text summarization approaches have been mostly based on metrics that measure similarities of system generated summaries with a set of human written gold-standard summaries (Hariharan and Srinivasan, 2010). Scientific article summarization is different from general domain summarization (e.g., Newswire data) which requires extensive analysis of Rouge’s effectiveness for scientific summarization. Works earlier have reported that Rouge is not much reliable in evaluating scientific summaries. An alternative metric for summarization evaluation which is based on the content relevance between a system generated summary and the corresponding human written summaries is proposed namely SERA (Summarization Evaluation by Relevance Analysis) which achieves high correlations with manual scores (Cohan and Goharian, 2016).

In order to cope with the growing number of relevant scientific publications to consider at a given time, automatic text summarization process is considered to be a powerful tool with several useful techniques. Summarizing these scientific papers poses important challenges for the natural language processing community (Saggion *et al*., 2016). In recent years, a number of evaluation challenges have been proposed to address the problem of summarizing a scientific paper taking advantage of its citation network (i.e., the papers that cite the given study). The researchers have presented a trainable technology to address a number of challenges in the context of the 2nd Computational Linguistics Scientific Document Summarization Shared Task.

Document summary generation faces instability in user agreement. Therefore, improvements in document summarization are necessary by analyzing the sentence position and recommendations of sentences from other sentences for generating good summary. As far scientific articles are concerned, challenge is very high. While human judgments are important in evaluating summaries, studies have proved the efficiency of the automated system much closer to user selection (manual summaries). This study focuses only on providing improvements for news articles. We have attempted to obtain summaries much closer or equal to manually generated summaries and the results obtained were promising. We also show that term frequency approach combined with position weight gives better results while adding node weight with the above feature yield results that were significantly better than the former approach. The study also illustrates some studies on some common evaluation criterion to generate a unique summary by various users. The results were also, compared with commercially existing Microsoft summarizer. The results produced by us were better as compared to the existing summarizers (Hariharan and Srinivasan, 2010).

MATERIALS AND METHODS

Characteristics of scientific documents: A scientific document is characterized by several important features in the document like: position, keyword and format. Each of these methods is discussed in this study to highlight the significance of each of these approaches.

Position based: Position-based scoring considers the sentence location. The most influential sentences are the sentences following the title-the first sentence of the text was included in the summary in 100% of the cases, the second and the third sentence in 65% of the cases. The sentences immediately following the subtitles were included in the 60% of the cases. The first sentence of the paragraph was included in the summary in 40% of the cases and the second and the third in 20% of the cases. In addition, 20% of the summaries contained the last sentence of the text. Table 1 presents a sample weightage awarded to the various positions of the sentences.

Table 1: Sample position based score

Feature	Percentage	Given score
1st sentence in article	100	10
2nd sentence in article	65	7
3rd sentence in article	65	7
1st sentence after sub heading	60	6
1st sentence in paragraph	40	4
2nd sentence in paragraph	20	2
Other	6	0

Table 2: Illustration for format based score

Feature	Percentage	Given score
Default	32	3
Bold of italic	70	10
Question or exclamation marks in sentence	10	0
Quotation marks in sentence	18	2
Captions, researchers, sub heading	0	0

Format based: Format-based scoring considers the sentence font (default, bold or italic) and punctuation marks (exclamation and question marks, double quotes) figure captions and the text researcher are also detected and given minimum scored. Table 2 presents a sample illustration of format based score.

Key word based: Key word-based scoring technique uses two techniques for identifying the key words. They are:

- Finding words that are relatively frequent in the study
- Not very frequent in general word frequency table extracting words from the text title and all subtitles

The words belonging to the title (article headline) and subtitles are given extra scores (5 and 2 points, respectively). All the other words are put into the local frequency table with a weight as.

RESULTS AND DISCUSSION

In this proposed framework, the summary of the technical study will be given in two formats-minimal summary and final summary. The minimal summary gives a brief overview of the input IEEE paper comprising the title, researcher name, abstract, table with selected rows, image and three recent references. In final summary, along with the minimal summary, the summarized text is included.

This process involves determining the weight of sentences by removing stop words, counting the number of occurrences of words in it and giving additional weight to the words that appear in the title. The sentences are then, ranked and assembled to form the final summary. The proposed research deals with summarization of the

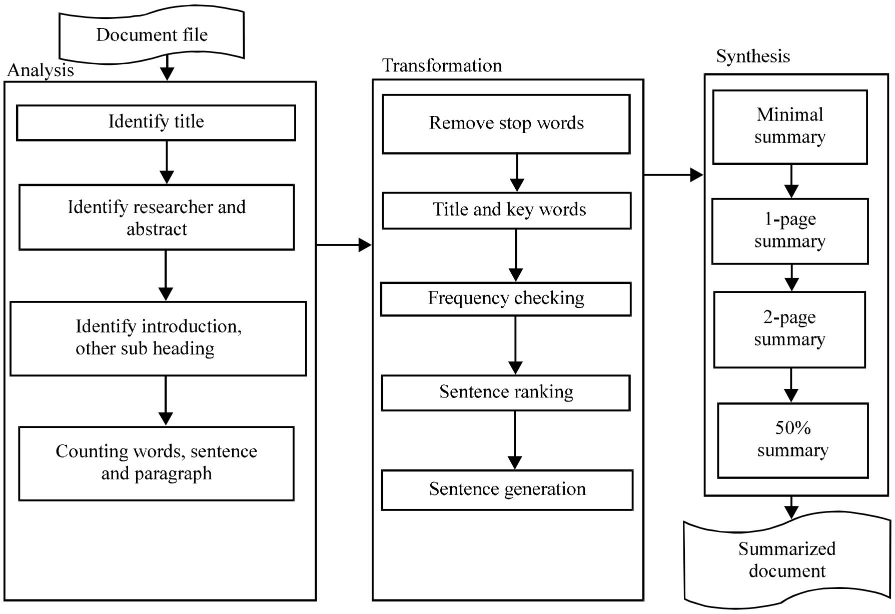


Fig. 1: Architecture of technical summarizer

Table 3: Special words for terms (based on location)

Criteria	Special weight
Title	4
Abstract	2
Heading	2
Sub-heading	2
Text body	1

input document (technical) which is an IEEE standard. The IEEE paper includes the title, author information, abstract, headings and sub headings, images, tables and references. The overall system architectural diagram given in Fig. 1.

The architecture involves three phases:

- Analysis
- Transformation
- Synthesis

In the analysis phase, we fragment the document content into title, researcher, paragraph headings and the body content adhering to the guidelines of IEEE paper format. Then, a statistical summary report of the number of words, pages, formatting styles used is generated. In the transformation phase, the stop words are removed (Table 3) and then, the stemming process (identifying the root word e.g., connect, connected, connection all three words give the rout word "connect) is done to add weightage to the keyword. Weightage is assigned to the words and sentence weightage is computed. Then, the sentences are ranked and finally, the summary is generated.

Sample list of stop words:

- "a"
- "I"
- "an"
- "at"
- "as"
- "in"
- "so"
- "if"
- "up"
- "he"
- "it's"
- "its"
- "and"
- "she"
- "not"
- "are"
- "and"
- "but"
- "was"
- "for"
- "have"
- "when"
- "what"
- "that"
- "then"

The weight assigned for a word is assigned by three cases. Consider there are 100 sentences. Then, summary generated is 25, 40 and 50% from the original summary list. Table 3 presents the arbitrary weights assigned for the

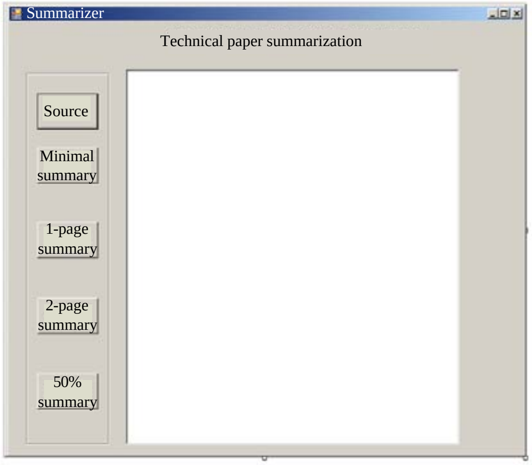
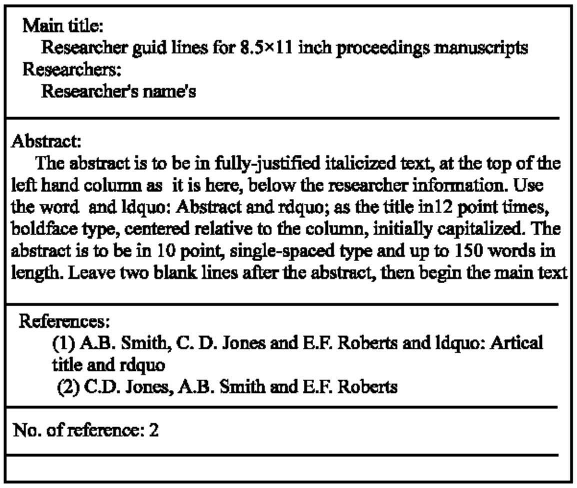


Fig. 2: Window showing minimal summary



Main title:
Researcher guid lines for 8.5×11 inch proceedings manuscripts
Researchers:
Researcher's name's

Abstract:
The abstract is to be in fully-justified italicized text, at the top of the left hand column as it is here, below the researcher information. Use the word and ldquo: Abstract and rdquo; as the title in12 point times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10 point, single-spaced type and up to 150 words in length. Leave two blank lines after the abstract, then begin the main text

References:
(1) A.B. Smith, C. D. Jones and E.F. Roberts and ldquo: Artical title and rdquo
(2) C.D. Jones, A.B. Smith and E.F. Roberts

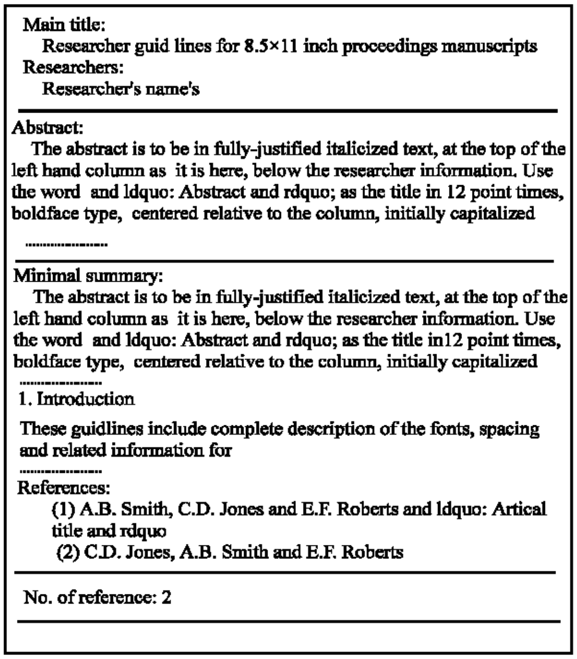
No. of reference: 2

Fig. 3: Window showing minimal summary

words based on its location in the document. These weights are added the along with the frequency of the words in the corresponding sentence.

The sorted sentences are arranged and depending upon the type of summary (user choice), the sentences were selected based on the percentage as noted in Table 4. For instance if the summary choice opted is 2-page, then 40% of sentences will be chosen for the output. Figure 2 presents the summary window for design for our proposed work. Figure 3-5 presents the minimal, one-page and 50% summary window, respectively. The output is categorized into 4 stages as minimal summary, 1-page summary, 2-page summary and 50% summaries. In minimal summary the work area displays the output in this format as:

- Title of the study
- Respective researchers



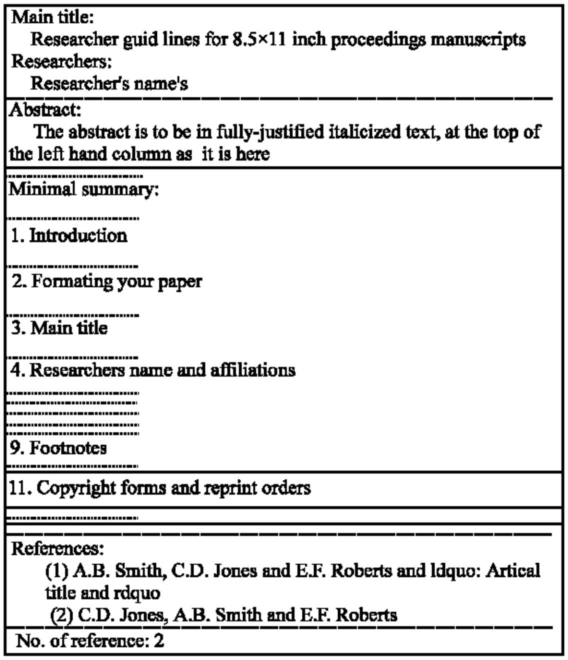
Main title:
Researcher guid lines for 8.5×11 inch proceedings manuscripts
Researchers:
Researcher's name's

Abstract:
The abstract is to be in fully-justified italicized text, at the top of the left hand column as it is here, below the researcher information. Use the word and ldquo: Abstract and rdquo; as the title in 12 point times, boldface type, centered relative to the column, initially capitalized
....

Minimal summary:
The abstract is to be in fully-justified italicized text, at the top of the left hand column as it is here, below the researcher information. Use the word and ldquo: Abstract and rdquo; as the title in12 point times, boldface type, centered relative to the column, initially capitalized
....
1. Introduction
These guidlines include complete description of the fonts, spacing and related information for
....
References:
(1) A.B. Smith, C.D. Jones and E.F. Roberts and ldquo: Artical title and rdquo
(2) C.D. Jones, A.B. Smith and E.F. Roberts

No. of reference: 2

Fig. 4: Window showing 1-page summary



Main title:
Researcher guid lines for 8.5×11 inch proceedings manuscripts
Researchers:
Researcher's name's
Abstract:
The abstract is to be in fully-justified italicized text, at the top of the left hand column as it is here
....
Minimal summary:
....
1. Introduction
....
2. Formating your paper
....
3. Main title
....
4. Researchers name and affiliations
....
....
....
....
9. Footnotes
....
11. Copyright forms and reprint orders
....
References:
(1) A.B. Smith, C.D. Jones and E.F. Roberts and ldquo: Artical title and rdquo
(2) C.D. Jones, A.B. Smith and E.F. Roberts
No. of reference: 2

Fig. 5: Window showing 50% summary window

Table 4: Summary type and percent of sentences selected

Summary type	Percentage
Minimal	20
1-page	25
2- page	40
50%	50
100%	100

- Abstract
- References

In 1-page summary the work area displays the output similar to minimal summary. In addition it includes the paragraph headings and sentences that meets the maximum number of sentences displayed in 1-page format.

In 2-page summary the work area displays the output same as 1-page report summary but with additional sentences that are included to meet the requirement of 2-page format. In 50% summary the work area displays the output same as the minimal summary and also includes additional paragraph headings and sub headings to meet with 50% stipulation.

CONCLUSION

The proposed researcher focuses on summarizing technical study in a standard format. The tool designed for the proposed task has been successfully designed, developed and tested. The framework proposed for the technical summarizer has been designed to handle document files of technical contents. It would also be able to handle many other file formats. Our system does not support any pictures, graphs and tables.

RECOMMENDATIONS

So, in future this project can be extended to include the pictures, graphs and tables also in final summary. Currently, summarizer handles only single document. We could extend the proposed task in experimenting with multiple document text contents and also in evaluating the generated summary.

REFERENCES

Baxendale, P., 1958. Machine-made index for technical literature-an experiment. IBM J. Res. Dev., 2: 354-361.

Cohan, A. and N. Goharian, 2016. Revisiting summarization evaluation for scienti?c articles. Proceedings of the 10th International Conference on Language Resources and Evaluation LREC, May 23-28, 2016, ELRA Publisher, Portoroz, Slovenia, pp: 251-263.

Edmundson, H.P., 1969. New methods in automatic extracting. J. Assoc. Comput. Machinery, 16: 264-285.

Filho, P.P.B. and T.A.S. Pardo, 2007. Summarizing scientific texts: Experiments with extractive summarizers. Proceedings of the 7th International Conference on Intelligent Systems Design and Applications ISDA, October 20-24, 2007, IEEE, Rio de Janeiro, Brazil, ISBN:978-0-7695-2976-9, pp: 520-524.

Hahn, U. and I. Mani, 2000. The challenges of automatic summarization. Comput., 33: 29-36.

Hariharan, S. and R. Srinivasan, 2010. Studies on intrinsic summary evaluation. Intl. J. Artif. Intell. Soft Comput., 2: 58-76.

Jbara, A.A. and D. Radev, 2011. Coherent citation-based summarization of scientific papers. Proceedings of the 49th Annual Conference on Human Language Technologies Vol. 1, June 19-24, 2011, Association for Computational Linguistics, Portland, Oregon, ISBN:978-1-932432-87-9, pp: 500-509.

Jing, H., 2000. Sentence reduction for automatic text summarization. Proceedings of the 6th Conference on Applied Natural Language Processing, April 29-May 04, 2000, Association for Computational Linguistics, Seattle, Washington, pp: 310-315.

Jung, Y., 2017. A semantic annotation framework for scientific publications. Qual. Quantity, 51: 1009-1025.

Luhn, H.P., 1958. The automatic creation of literature abstracts. IBM J. Res. Dev., 2: 159-165.

Mallett, D., J. Elding and M.A. Nascimento, 2004. Information-content based sentence extraction for text summarization. Proceedings of the International Conference on Information Technology: Coding and Computing ITCC Vol. 2, April 5-7, 2004, IEEE, Las Vegas, Nevada, USA., ISBN:0-7695-2108-8, pp: 214-218.

Mohamed, S.S. and S. Hariharan, 2016. Parameters affecting the judgment task while summarizing documents. Intl. Arab J. Inf. Technol., 13: 417-426.

Mohd, M., M.B. Shah, S.A. Bhat, U.B. Kawa and H.A. Khanday *et al.*, 2016. Sumdoc: A Unified Approach for Automatic Text Summarization. In: Proceedings of the 5th International Conference on Soft Computing for Problem Solving, Pant, M., K. Deep, J. Bansal, A. Nagar and K. Das (Eds.). Springer, Singapore, ISBN:978-981-10-0447-6, pp: 333-343.

Muurisep, K. and P. Mutso, 2005. ESTSUM-Estonian newspaper texts summarizer. Proceedings of the 2nd Baltic Conference on Human Language Technologies, April 4-5, 2005, University of Tartu, Tartu, Estonia, ISBN: 9985894839, pp: 311-316.

Patil, V., M. Krishnamoorthy, P. Oke and M. Kiruthika, 2013. A statistical approach for document summarization. Intl. J. Adv. Comput. Technol., 2: 33-43.

Qazvinian, V. and D.R. Radev, 2008. Scientific paper summarization using citation summary networks. Proceedings of the 22nd International Conference on Computational Linguistics Vol. 1, August 18-22, 2008, Association for Computational Linguistics, Manchester, England, ISBN:978-1-905593-44-6, pp: 689-696.

Qazvinian, V., D.R. Radev, S.M. Mohammad, B. Dorr and D. Zajic *et al*., 2013. Generating extractive summaries of scientific paradigms. J. Artif. Intell. Res., 46: 165-201.

Saggion, H., A. AbuRa'ed and F. Ronzano, 2016. Trainable citation-enhanced summarization of scientific articles. Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries BIRNDL, June 23, 2016, SIG Publiser, Newark, New Jersey, pp: 175-186.

Tabassum, S. and E. Oliveira, 2015. A review of recent progress in multi document summarization. Proceedings of the 10th Doctoral Symposium in Informatics Engineering DSIE'15, January 29-30, 2015, University of Porto, Porto, Portugal, ISBN:978-972-752-173-9, pp: 48-59.

Teufel, S. and M. Moens, 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. Computat. Ling., 28: 409-445.