ijar
ISSN NO. 2320-5407

Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/7031
DOI URL: http://dx.doi.org/10.21474/IJAR01/7031

RESEARCH ARTICLE

DOCUMENT SUMMARIZATION USING SENTENCE BASED TOPIC MODELING AND CLUSTERING

Augustine George[1] and Dr. Hanumanthappa[2].

1. Research Scholar, Bharathiyar University.
2. Professor, Bangalore University.

Manuscript Info

Manuscript History

Received: 04 March 2018
Final Accepted: 06 April 2018
Published: May 2018

Keywords: -
Term Frequency, Natural Language Processing, Text Summarization, Structural Topic Modeling (STM), Pre-processing, Tokenization, Topic Modeling**.**

Abstract

In recent years, the practical application of automatic document summarization has become popular and numerous papers published based on the topic. There are many approaches to identify the significant portion of each document. Topic representation and modelling is an intermediate representation of the text that captures the topics discussed in the input and aids the automatic summarization. The significance of sentences decided based on the representations of topics in the input document. This article attempts to provide a comprehensive summary that includes sentence extraction, tokenization on the extracted sentences. Sentence based Structural Topic Modeling (STM) is used to determine important content for each domain in the integrated document and sentences are grouped using k-means clustering under each topic. Further Text Summarization of sentences under each topic achieved using its Term Frequency of each sentence. Finally**,** the sentences are arranged based on its Lexical Ranking score in the summarized text.

Introduction: -

Text abstraction has assumed great importance in an era where the emphasis is on multi-document and multilingual summarization. Most summarizers generate summaries from multiple documents by some kind of an extraction algorithm and then present the concatenated content to the reader [1]. But it is also indispensable that they must be semantically meaningful to enhance the comprehensiveness of the summary being generated. A lot of research focused on language processing and machine-learning techniques to determine which sentences could better represent the original text. Significant advances were made in determining the semantic interpretations and developing similarity measures. Researchers are more interested to explore new models for summarization and investigating a variety of approaches to come up with accurate summarization using Natural Language Processing (NLP). Topics give a high-level description of its contents to the reader, and hence we have used a hybrid approach to tackle the problems of summarization using the STM package.

Literature Survey: -

In the recent past, several methods proposed based on statistical, graph-based, and machine learning approaches. Statistical Approaches focus on feature extraction that the sentences selected based on concepts such as co-occurrences of words and classification.

Author [2] proposed Bayesian classification in text summarization; the system defined for single document

Corresponding Author:- Augustine George.
Address:- Research Scholar, Bharathiyar University.

summarization (SDS) in Japanese and, in some cases, evaluated more accurately than other systems. There are also researches [3] that used Lexical chains and Bayesian method for sentence extraction. Statistical approaches fail for new corpora from a different domain. Graph-based approach performs well topic-driven summarization by obtaining a sub-graph that is similar to the topic. Summarization can be done by identifying most connected node from each subgraph. It is popular in ranking web pages by Google's PageRank [7] that serves an index and search for relevant web pages. For example, systems like TextRank [4], LexRank [5], and Hypersum [6] use graph-based approach to obtain good summaries but less significant for multi-document summarization.

Author/Year	**Technique**	**Text Representation**	**Content Selection**	**Summary Generation**
Barzilay and McKeown, 2005	Tree based	Dependency tree	An algorithm uses local alignment across a pair of parsed sentences	Algorithm for altering phrases from input sentences.
Greenbacker, 2011	Multimodal semantic model	Semantic model	Information Density	Synchronous tree
Genest and Lapalme, 2012	Rule-Based	Categorical and parametric	Rules	Generation patterns
Moawad and Aref 2012	Semantic graph	Rich semantic graph	C_{weight} =Average weight of each concept. $_{\text{The sweight}}$=average weight of all concepts in a sentence.	Reduced semantic graph and domain ontology

The proposed approach of summarization consists of the following two phases: ***Topic identification using Structural Topic Modeling (STM), Clutering of Sentences under each Topic using Centroid Based Clustering and arranging the sentences under each topic using Lexical Ranking***. The rest of the paper is organized as follows. Section 3 describes different phases of text summarization. Section 4 suggests the discussions on the obtained values and Section 5 gives conclusions and directions for future work.

Methodology: -

Most of the summarization mechanisms generate output in four stages; pre-processing, evaluations, information selection, output generation. The pre-processing stage involves removing meta-data, titles, figures from the original corpora for further analysis and the evaluation of information. The evaluation stage involves the analysis of datasets using machine learning algorithms to get the information required for the next steps.

In our previous work we used extractive method to integrate multiple documents and extracted relevant abstract using iterative elimination of matrix subsets and cosine similarity. We identified a method of integrating multi-document by retaining unique sentences from all the documents and removed repeated sentences using cosine similarity of different documents.

Sentences are reviewed and selected using sentence clustering algorithms in the information selection stage; the selected sentences are put together for output generation. The hypothesis of this paper is that a text summarization methodology using NLP techniques that provide optimized summary results regarding quality and scalability.

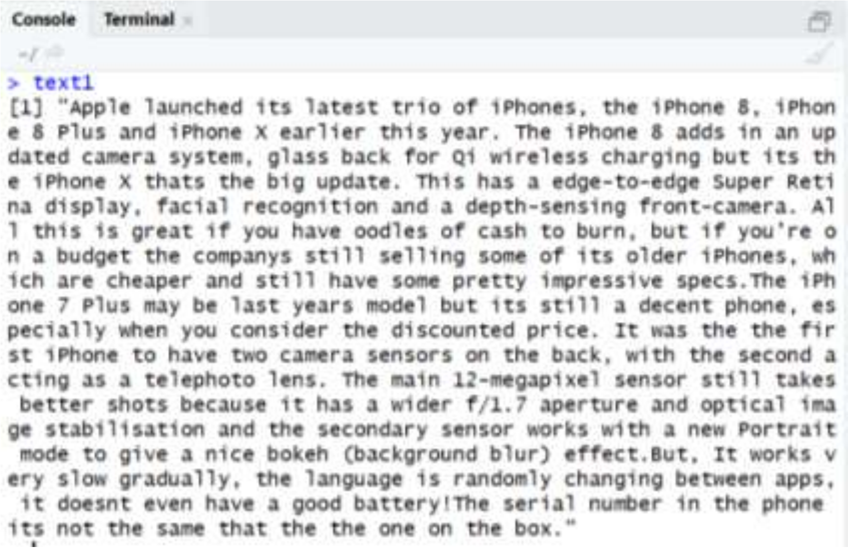
Steps

1. Tokenisation
2. Structural Topic Modeling
3. Centroid Based Clustering
4. Lexical Ranking

Tokenisation:-

A token is “linguistically significant and methodically useful”[10]. Identifying the sentence boundaries and tokenizing them into a data frame makes manipulations easier.

Sample data frame after pre-processing is shown in figure 1.



```
> text1
[1] "Apple launched its latest trio of iPhones, the iPhone 8, iPhon
e 8 Plus and iPhone X earlier this year. The iPhone 8 adds in an up
dated camera system, glass back for Qi wireless charging but its th
e iPhone X thats the big update. This has a edge-to-edge Super Reti
na display, facial recognition and a depth-sensing front-camera. Al
l this is great if you have oodles of cash to burn, but if you're o
n a budget the companys still selling some of its older iPhones, wh
ich are cheaper and still have some pretty impressive specs.The iPh
one 7 Plus may be last years model but its still a decent phone, es
pecially when you consider the discounted price. It was the the fir
st iPhone to have two camera sensors on the back, with the second a
cting as a telephoto lens. The main 12-megapixel sensor still takes
 better shots because it has a wider f/1.7 aperture and optical ima
ge stabilisation and the secondary sensor works with a new Portrait
 mode to give a nice bokeh (background blur) effect.But, It works v
ery slow gradually, the language is randomly changing between apps,
 it doesnt even have a good battery!The serial number in the phone
its not the same that the the one on the box."
```

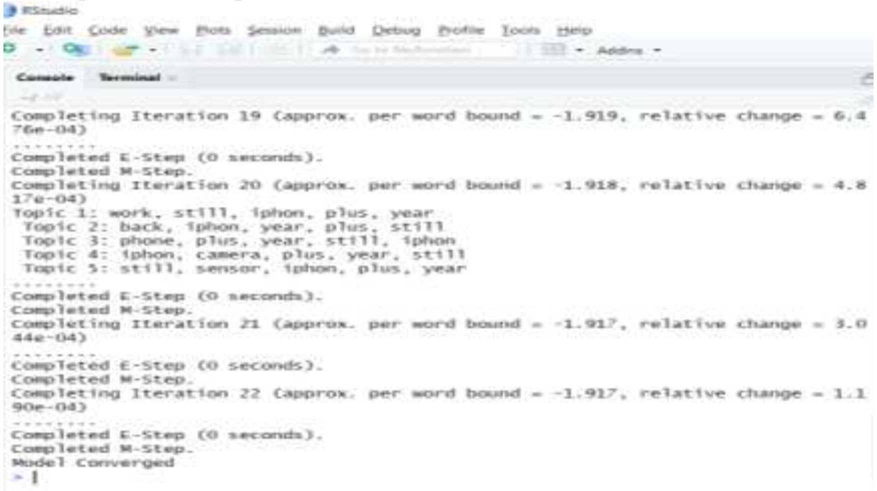
Figure 1:- Pre-processed Data frame

Structural Topic Modeling: -

Topic modelling technique is applied to the integrated individual text document cluster to generate the cluster topics and terms belonging to each cluster topic.

Structural topic models are a recent innovation in the area of multi-document summarization that enables the summarizer to improve identification of latent semantics in unstructured text data using metadata that describes different properties of a text. The STM package of R Software allows exploring different topics on the given documents, estimating the uncertainty of each of the topic, and visualizing attributes of interest of each topic. Structural Topic Modeling uses LDA (Latent Dirichlet Allocation) method that automatically discovers topics that the sentences contain. Latent Dirichlet Allocation (LDA) [11] is a popular topic modelling technique which models text documents as mixtures of latent topics that identify key concepts presented in the text. A topic model is a probability distribution technique where each document is modelled as a combination of topics that also represents groups of words that tend to occur together. At this stage, frequent terms are generated from the collection of the integrated text document. For example, given these sentences and asked for two topics, LDA might produce results, for instance, is represented in Figure 2.

- Sentences 1 and 2: 100% Topic A
- Sentences 3 and 4: 100% Topic B
- Sentence 5: 60% Topic A, 40% Topic B


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Console
Terminal

```
Completing Iteration 19 (approx. per word bound = -1.919, relative change = 6.4
76e-04)
........
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 20 (approx. per word bound = -1.918, relative change = 4.8
17e-04)
Topic 1: work, still, iphon, plus, year
 Topic 2: back, iphon, year, plus, still
 Topic 3: phone, plus, year, still, iphon
 Topic 4: iphon, camera, plus, year, still
 Topic 5: still, sensor, iphon, plus, year
........
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 21 (approx. per word bound = -1.917, relative change = 3.0
44e-04)
........
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 22 (approx. per word bound = -1.917, relative change = 1.1
90e-04)
........
Completed E-Step (0 seconds).
Completed M-Step.
Model Converged
> |
```

Figure 1:- instance of data frame after preprocessing

Clustering of Sentences based on topics: -

Clustering ensures that similar set of sentences having similar meaning or semantic are grouped together and logically represents a topic for effective summarization. The impact of clustering for summarization of large text collection is also demonstrated in this research paper. It is shown that summarization of topics using clustering gives better summarization performance as compared to the summarization without clustering.

Each topic is a distribution of words. Each document is a mixture of topics, and each word is drawn from those topics. Most often, while dealing with multiple documents, only the title is observed. The other contents might not be highlighted. Therefore, we use topic modelling to infer these hidden attributes. But mostly when topic modelling is done, the user has no idea how many topics are present. Therefore, the value of ‘k' is randomly assigned. Therefore, to just extract only the high level concepts and main keywords from the text, we take only the top topics that are shown in Figure 3 and Figure 4.

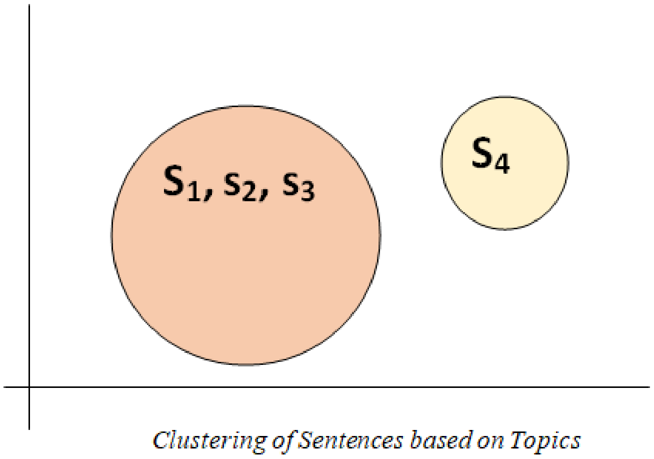


Figure 3:- Plotting the clusters of sentences based on the themes/ Topics

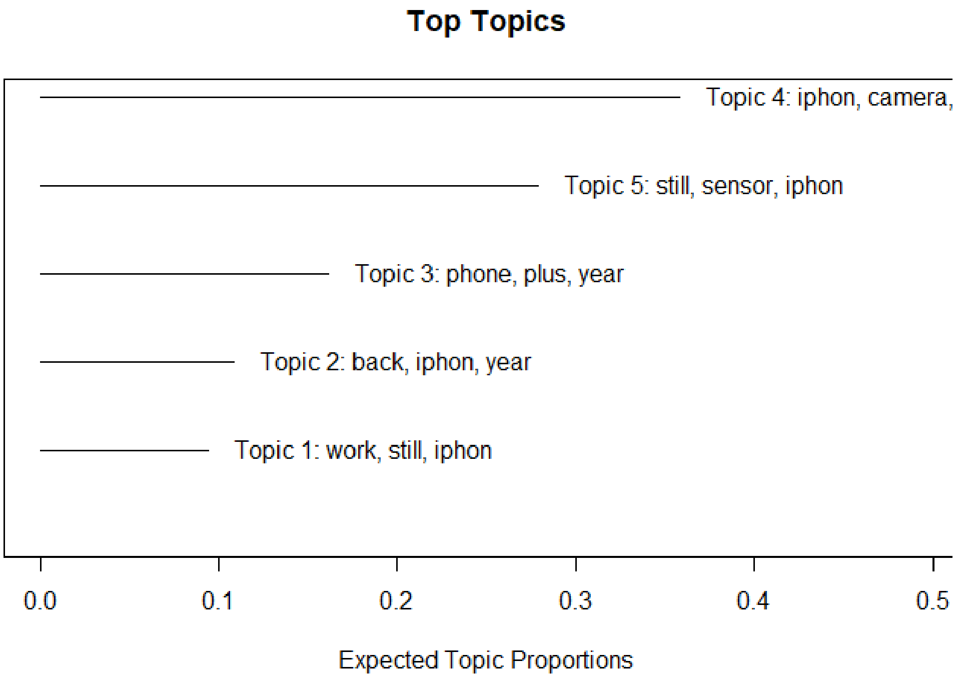
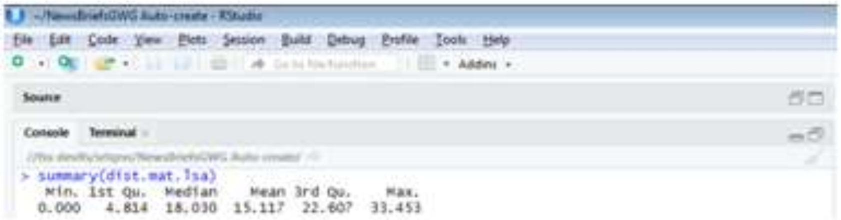


Figure 4:- Topic Proportions

Centroid-based clustering groups the set of sentences under each topic. The iterative clustering ends when all the sentences are grouped under the topics whose centroid cohesion is greater than others. Topics that are having less centroid cohesion will be eliminated automatically that shows no sentences are grouped under that particular topic. Sample centroid cohesion is shown in the following section.

a. Sentence to centroid cohesion:-

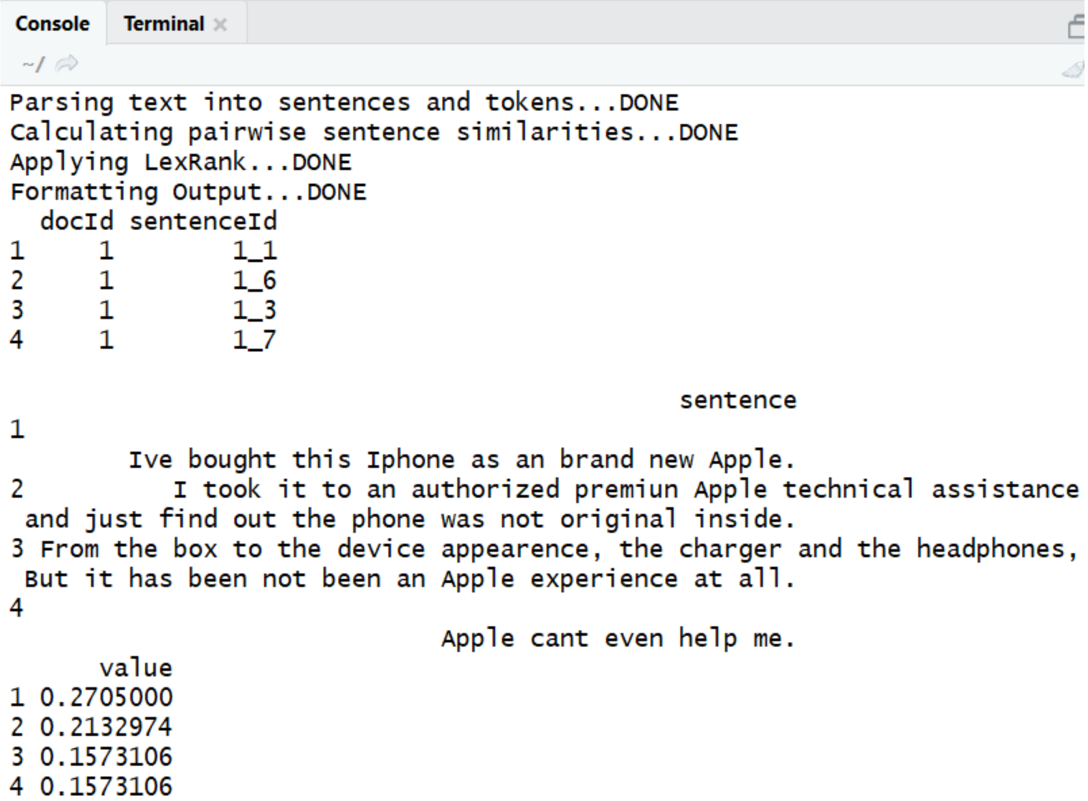
The centrality of a sentence is often defined regarding the centrality of the words that it contains. First, we compute the vector representing the centroid of the document, which is the arithmetic average over the corresponding coordinate values of all the sentences of the document; then we only extract the centroid sentence from each cluster. The centroid of a cluster is a pseudo-document which consists of words that have tf×idf scores above a predefined threshold. Sample centroid cohesion is shown in Figure 5.


~/NewsBriefsGWG Auto-create - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Source
Console Terminal

```
> summary(dist.mat.lsa)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   4.814  18.030  15.117  22.607  33.453
```

Figure 5:- Selecting the central values from a given set of topics; Eliminating correlated themes.

After grouping the relevant sentences under each topic, a lexical ranking is applied on all the sentences to ensure the association between lexical units which have different relative values based on some shared underlying semantic property. Author [12] identified N-Gram based linguistic processing approach that predicts the preceding word knowing the previous content with the probability of a word sequence in a newspaper article. LexRank applies a heuristic post-processing step on the summarization that builds up a summary by adding sentences in rank order but discards any sentences that are too similar to ones already placed in the summary. The method used is called Cross-Sentence Information Subsumption (CSIS). The output of lexrank and summary is shown in Figure 6.



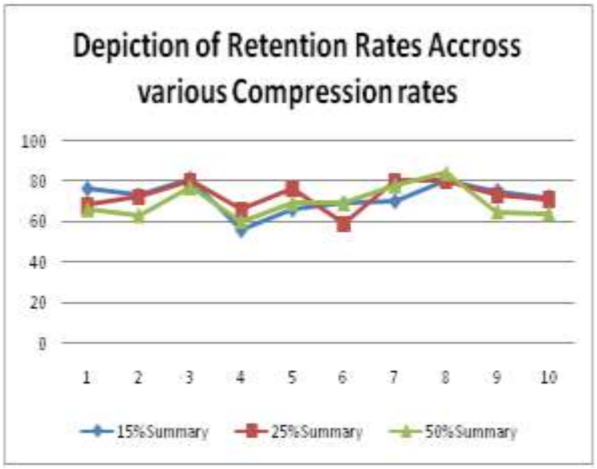
```
Parsing text into sentences and tokens...DONE
Calculating pairwise sentence similarities...DONE
Applying LexRank...DONE
Formatting Output...DONE
  docId sentenceId
1     1        1_1
2     1        1_6
3     1        1_3
4     1        1_7

                                             sentence
1
        Ive bought this Iphone as an brand new Apple.
2          I took it to an authorized premiun Apple technical assistance
 and just find out the phone was not original inside.
3 From the box to the device appearence, the charger and the headphones,
 But it has been not been an Apple experience at all.
4
                             Apple cant even help me.
      value
1 0.2705000
2 0.2132974
3 0.1573106
4 0.1573106
```

Figure 6:- Lexrank and Output summary.

Experimented Results and Discussion:-

After summarization, we examined the retention rate of sentences with an effect of different compression rates and depicted in the following Figure 7. **There is no considerable difference observed. The Highest RR achieved is 84% and the lowest being 56%.**



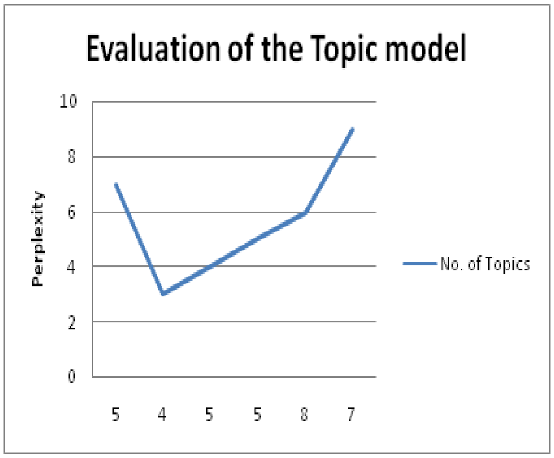


Fig. 7:- the Retention rate of sentences in the summary

Figure 8:- Perplexity of topics

Perplexity showed in the above figure depicts how well the proposed model predicts expected number of topics. Improvement of the proposed model would be to evaluate the created summary from a global perspective. The algorithm currently facilitates the check of global cohesion only in the theme construction phase, where the topics are spread evenly in some clusters. This guarantees that the final summary will cover all the important issues.

Conclusion: -

This paper presented document summarization technique using STM and clustering. It mainly works at the sentence level, focusing on the semantical relations between concepts. The presented technique is evaluated regarding scalability and various text summarization parameters namely; compression ratio, retention ratio. ROUGE and Pyramid score will also be measured in our future work. We also plan to investigate how to generate coherent temporally based event summaries. We will also investigate how users can effectively use multi-document summarization through interactive interfaces to browse and explore large document sets.

References: -

1. Ravikiran Vadlapudi , Rahul Katragadda "On Automated Evaluation of readability of summaries: Capturing Grammaticality, Focus, Structure and coherence.", Proceedings of the NAACL HLT 2010 Student Research Workshop, pages 7–12, Los Angeles, California, June 2010. C 2010 Association for Computational Linguistics.
2. Tadashi Nomoto. Bayesian learning in text summarization. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT ’05, pages 249–256, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics
3. Hal Daume, III and Daniel Marcu. Bayesian query-focused summarization. In ´ Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44, pages 305–312, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics
4. R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In Proceedings of EMNLP04and the 2004 Conference on Empirical Methods in Natural Language Processing, July 2004.
5. Tadashi Nomoto. Bayesian learning in text summarization. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT ’05, pages 249–256, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
6. Wei Wang, Furu Wei, Wenjie Li, and Sujian Li. Hypersum: hypergraph-based semisupervised sentence ranking for query-oriented summarization. In Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, pages 1855–1858, New York, NY, USA, 2009. ACM.

7. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120
8. Barzilay and McKeown, 2005, Sentence Fusion for Multidocument News Summarization, Journal of Computational Linguistics, Vol 31, issue 3, PP 297-328
9. Greenbacker, 2011, Towards a Framework for Abstractive Summarization of Multimodal Documents, Proceedings of the ACL-HLT 2011 Student Session, pages 75–80, Portland, OR, USA 19-24 June 2011. Association for Computational Linguistics
10. Genest and Lapalme, 2012, Fully Abstractive Approach to Guided Summarization, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 354–358, Jeju, Republic of Korea, 8-14 July 2012. c 2012 Association for Computational Linguistics
11. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. The Journal of Machine Learning Research 3:993–1022
12. Deepa Nagalavi, M.Hanumanthappa, N-gram Word prediction language models to identify the sequence of article blocks in English e-newspapers, Proceedings of International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), IEEE, 10.1109/CSITSS.2016.7779376 ISBN: 978-1-5090-1022-6.