ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Automatic Text Summarization Using Regression Model (GA)

Anil Kumar, JyotiYadav, Seema Rani

M.Tech Student, Dept. of CE, YMCA University of Science & Technology, Faridabad, India

ABSTRACT: Text Summarization provide large text data into a shorter version without changing its information content and meaning. It is very difficult for human beings to read whole document to understand and manually summarize the documents. Text Summarization methods are divide into two types: extractive and abstractive summarization. Anabstractive summarization is understanding of document, finding the new concepts and providing summary in few words (sentences different form texts sentences).it is very hard or impossible to design (now a days). In Extractive summarization method select important sentences, paragraphs etc. from the original text document and concatenating them into shorter form. The importance of sentences is decided based on some features of sentences. In this Research paper, Automatic Text Summarization using Extractive techniques with the help of Genetic Algorithm has been presented.

KEYWORDS: Automated Text Summarization; extract relevant, non-redundant.

I. Introduction

Now a day's size of document is very large, it is difficult to read or understand whole document so, summary of document is needed without changing the main content of document.It is time consuming and very difficult task to extract information manually from large amount of data available. Therefore, there is a need to extract relevant, non-redundant and important data. By Automated Text Summarization (ATS) this tasks can be made easier. ATS automatically convert big text file into summarized form, which is easily understandable, readable and complete. Without ATS, it is very laborious job for human to read out whole Document to understand. An example, let's suppose any person want a very short summary of every email message on his mobile phone, so by reading that summary he may take decision that he has to pay attention to that mail instantly or not [9] and some other areas where ATS can be applied are Media, Medical, Education, Curriculum Vitae, Sports, Preview of moviesand Subject of an email or letter etc. [1][2].

II. RELATED WORK

Text summarization is very important because it describe a large document in few lines without changing the original meaning of document. Selection of important lines is the main problem. Many research scholars have explored various method or features.

P.B. Baxendale [3] in 1958 proposed new feature that is sentence position or sentence location in an input document. Baxendaleanalyzed that sentences which are located at the beginning or at the end of the document are important than other sentences contained into the document. He tested sentence location feature on 200 paragraphs and found that 85% of the paragraphs were from beginning and 7% of the paragraphs were from ending locations. A sentence location feature became an important feature for sentence extraction and it is used till now.

D.R. Radev et al [4] in 2004 developed a system for multi-document summarization named centroid-based summarization (CBS). The first phase was topic detection, here topic related to same event were grouped together. An agglomerative algorithm was used over TF-IDF for accomplishing this task of topic detection. In second phase centroids values were used to identify sentences in every cluster that were central to the topic of whole cluster. Here two metric were defined by author that were used to resemble the two in the maximal marginal relevance (MMR). The first metric was cluster based relative utility (CBRU), which evaluate the relevance of a particular sentence to the general topic of whole cluster. The second metric was cross-sentence informational subsumption (CSIS), which compute redundancy factor among sentences. Three features were used for calculating score of every sentence Si. Centroid value, Sentence Position, First-sentence overlap.H.P. Edmundson [5] in 1969 described new typical structure for a text summarization. He proposed two new features and incorporates two old features explained above. Two new features proposed by Edmundson are: Cue Words, Title or Heading Words.

III.PROPOSED WORK

Usually, the information in a given document is not constant, which means that some parts of document are more important than others are less important. The main challenge is to identify important parts of document and extract them for final summary. Here most work presented on single-document summarization using extraction method. In this section, some extractive techniques are discussed briefly, which are applied for extraction of sentences for final summary. Extraction technique is divided into two steps1. Pre Processing 2. Processing. Preprocessing phase involves three steps a) Sentences boundary identification. b) Stop-Word Elimination c) Stemming.In processing phase, feature value for every sentence is calculated. Score of every sentence between 0 to 1 and then weights are assigned to these features using weight learning method.

A. TEXT FEATURES

1. Sentence Position (f1)

We assume that the first sentences of a paragraph are the most important. Therefore, we rank a paragraph sentence according to their position and we consider maximum positions of 5. For instance, the first sentence in a paragraph has a score value of 5/5, the second sentence has a score 4/5, and so on.

2. Positive keywords (f2)

Positive keywords in the sentence are the keywords come many times in the summary. It is calculated as formula given in fig. No 1

$$Score_{f2}(s) = \frac{1}{Length(s)} \sum_{i=1}^{n} tf_i * P(s \in S|keyword_i)$$

$$P(s \in S|keyword_i) = \frac{P(keyword_i|s \in S)P(s \in S)}{P(keyword_i)}$$

$$P(keyword_i|s \in S) = \frac{\#(Sentence\ in\ summary, and\ contains\ keyword_i)}{\#(Sentence\ in\ summary)}$$

$$P(s \in S) = \frac{\#(Sentence\ in\ training\ corpus, and\ also\ in\ summary)}{\#(Sentence\ in\ training\ corpus)}$$

$$P(keyword_i|) = \frac{\#(Sentence\ in\ training\ corpus\ , and\ contains\ keyword_i)}{\#(Sentence\ in\ training\ corpus)}$$

Where, s is a sentence, n is the number of keywords in s,
tfi is the occurring frequency of keywordi in s.
Divide the value by the sentence length to avoid the bias of its length

Fig .1 Formula for Positive Keywords.

3. Negative keywords (f3)

Negative keywords are the keywords that are unlikely to occur in the summary. It can be calculated as formula given in fig. No 2

$$Score_{f3}(s) = \frac{1}{Length(s)} \sum_{i=1}^{n} tf_i * P(s \in S|keyword_i)$$

Fig .2 Formula for Negative Keywords.

4. Sentence Centrality (f4)

Sentence centrality is the vocabulary overlap between this sentence and other sentences in the document. It is calculated as formula given in fig. No 3

$$\mathrm{Score}_{f4}(s) = \left| \frac{\text{Keywords in s} \cap \text{Keywords in other Sentences}}{\text{Keywords in s} \cup \text{Keywords in other Sentences}} \right|$$

Fig .3 Formula for Sentence Centrality.

5. Numerical data (f5)

The sentences that contains numerical data is an important one and it is most probably included in the document summary. It is calculated as formula given in fig. No 4

$$Score_{f5}(s) = \frac{\#(numerical\ data\ in\ s)}{Length(s)}$$

Where, Num (s) is numerical data present in sentence s
Length(s) is length of sentence s

Fig .4 Formula for Numerical Data.

6. Presence of Brackets (f6)

Sometimes sentences may contain brackets such as ( ) parentheses. Mostly braces contains material which could be omitted without destroying or altering sentence meaning. After doing analysis it has been found that brackets do not contain important information and has lower probability to be included for the summary[9]. Presence of brackets in sentence is calculated as formula given in fig. No 5

$$Score_{f6}(s) = \frac{SenLen(S) - BracLen(S)}{SenLen(S)}$$

Where, SenLen(S) Sentence length of a sentence
BracLen(s) Brackets length in a sentence

Fig .5 Formula for find Brackets.

7. Presence of inverted Commas(f7)

In texts (“ ”) double quotation marks or inverted comma surrounding quotations, direct speech, literal title or name etc. contains important information. After doing analysis it has been found that an inverted comma has higher probability to be included for the summary. Presence of inverted commas is calculated as formula given in fig. No 6

$$Score_{f7}(s) = \frac{QuoteWords(S)}{SenLen(S)}$$

Where, SenLen(S) Sentence length of a sentence
QuoteWords(s) are the words present in Quote’s

Fig .6 Formula for find inverted commas.

8. Sentence Length (f8)

Sentences which are shorter in length may not represent theme of a text document because of fewer words contained in it, although selecting longer length sentences are also not good for summary. So sentence length values are calculated in such a way that, shorter and longer sentences are assigned lower values. Sentence length values are calculated as formula given in fig. No 7

$$Score_{f8}(s) = \frac{Words(Si)}{SenLen(Si)}$$

Fig .7 Formula for Sentence Length.

9. Presence of Commas(f9)

Sentences containing common words are important. These words having more information and have higher probability to be extracted for the summary. Common words are calculated as formula given in fig. No 8

$$Score_{f9}(s) = \frac{\#\,Comma(Si)}{SenLen(Si)}$$

Fig .8 Formula for calculating Commas.

10. Presence of Acronym(f10)

Sentence contains Acronym words are the most important sentence in a paragraph which are written in capital letters. Eg. YMCA. Acronym Words are Calculated as formula given in fig. No 9

$$Score_{f10}(s) = \frac{\#\,Swords(Si)}{SenLen(Si)}$$

Where, Swords are the Acronym words

Fig .9 Formula for Acronym word.

B. **GENETIC ALGORITHM MODEL (GA):-**Genetic Algorithms are a way of solving problems by mimicking the same processes Mother Nature uses. Therefore, GA can be used to specify the weight of each text feature. For a sentence s, a weighted score function, as shown in above given formulas (f1 to f10) is exploited to integrate all the ten feature scores, here wi indicates the weight of fi.[6][7]
In GA, training mode is used for trained a model and that model (learned feature weights) fig.10 is used in testing mode for generate summary.

Training Mode:-

In Training Mode features are extracted from the manually summarized documents and after apply GA feature weights are generated.

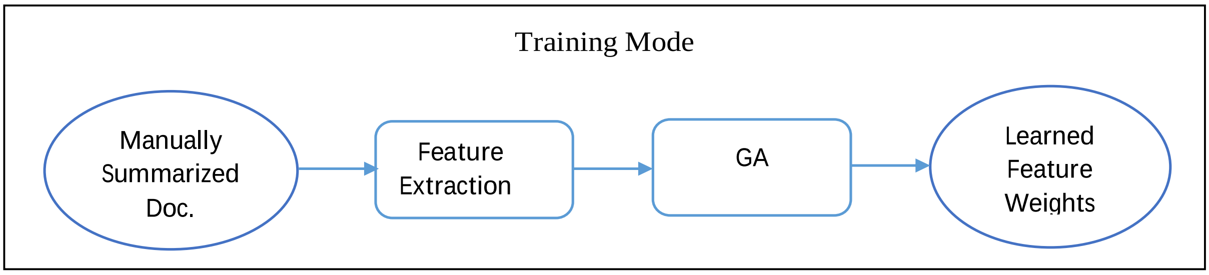


Fig .10Training Mode

Testing Mode:-Feature weights are generated in training mode and these weights are used form generate the summary of different documents in Testing Mode.

DOI: 10.15680/ijircce.2015.0305123

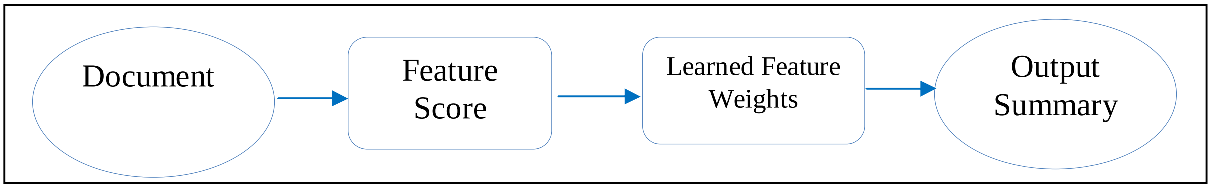


Fig .11Testing Mode

Sentence Scoring Phase

In Sentence Scoring phase firstly weights of every sentence is calculated, Then Score of every sentence is calculated. After calculating final score of every sentence, extraction of sentences is done according to compression ratio required.

$$\text{Score(s)} = w_1.Score_{f1}(\text{s}) + w_2.Score_{f2}(\text{s}) + w_3.Score_{f3}(\text{s}) + w_4.Score_{f4}(\text{s}) + w_5.Score_{f5}(\text{s}) + w_6.Score_{f6}(\text{s}) + w_7.Score_{f7}(\text{s}) + w_8.Score_{f8}(\text{s}) + w_9.Score_{f9}(\text{s}) + w_{10}.Score_{f10}(\text{s})$$

Fig .12 Formula for finding Sentence Score

In this 10 different text features are used to score sentences. After each sentence of a document is scored, the sentences of the document are sorted according to their scores and the highest scored sentences are selected to form the summary of that document. However, not all the feature scores have the same importance while calculating the sentence score. A sentence score is a weighted sum of that sentence's feature scores. Each feature may have a different weight and these weights are learned from the manually summarized documents, using machine learning methods. Thus, a sentence's score is calculated.

fi are the feature scores of each sentence and their values can range from 0 to 1.They are computed separately for each sentence s. wi can range from 11 to 25. They are learned using genetic algorithms. Here, which feature is used many times or for important feature weight(wi) is high and so on.

Chromosome of Weights

Chromosome are the n number of sets of 10 weights.A chromosome is represented as the combination of all feature weights. For every sentence chromosomes are different. These are used for find set of weights on which a sentence give highest sentence score.

Wi=

w11,w12,w13,w14,w15,w16,w17,w18,w19,w110
w21,w22w23,w24,w25,w26,w27,w28,w29,w210
w31,w32,w33,w34,w35,w36,w37,w38,w39,w310
w41,w42,w43,w44,w45,w46,w47,w48,w49,w410
- - - - - - - - - -
- - - - - - - - - -
wn1,wn2,wn3,wn4,wn5,wn6,wn7,wn8,wn9,wn10

Fig .13Chromosome of weights

Finding Best Weights from Training Mode

After finding the chromosome for every sentence a set of best weights is generated (on which sentence gives highest score) for every sentence. When weights for every sentence is find out then these weights cross-over with every sentence feature. After cross-over a single best one weight is generated i.e. Wi1.

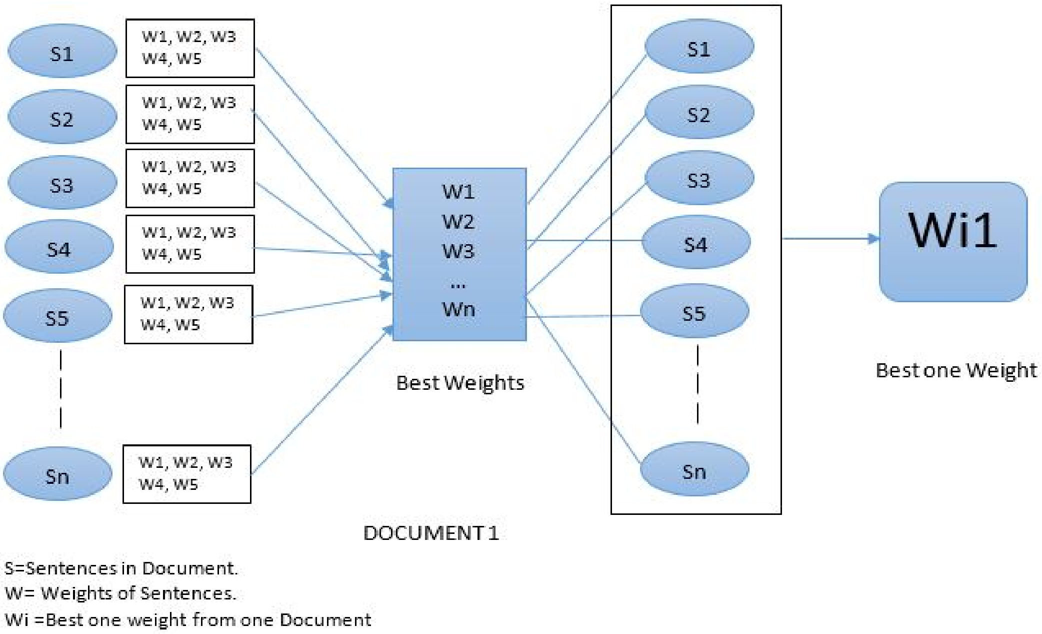


Fig .14 Finding Best Weight from one document.

Best one weight is generated for every documenton the basis of highest score of sentences (fig.14)and these weights are apply on every document sentences. These weights are apply according to weights rank equation (fig.13). Scores of every sentences are generated and one set of weights are selected on which maximum number of highest sentence scores are generated. There is one best score weight is generated and this weight is used for further summarized any text document.

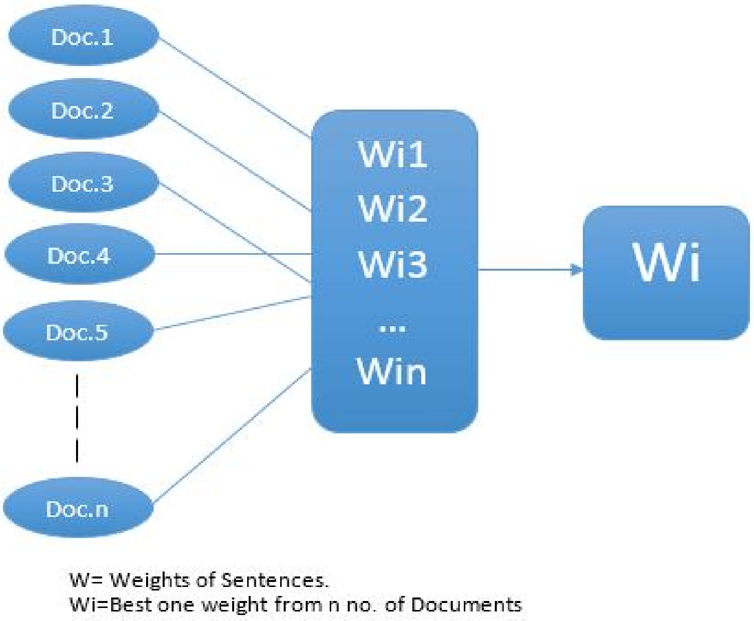


Fig .15 Finding Best Weight for n no. of document.

In order to make summary more reliable, accurate, complete and less redundant following process is applied:
1. Final weight of every sentence using weight-ranking equation is computed.
2. Final weights are sorted in reverse order.
3. Best one weight is used for generating summary.
4. Top weighted sentences are selected for summary according to compression ratio required.
5. Selected sentences for summary are shown in same sequence as they appeared in input text document.

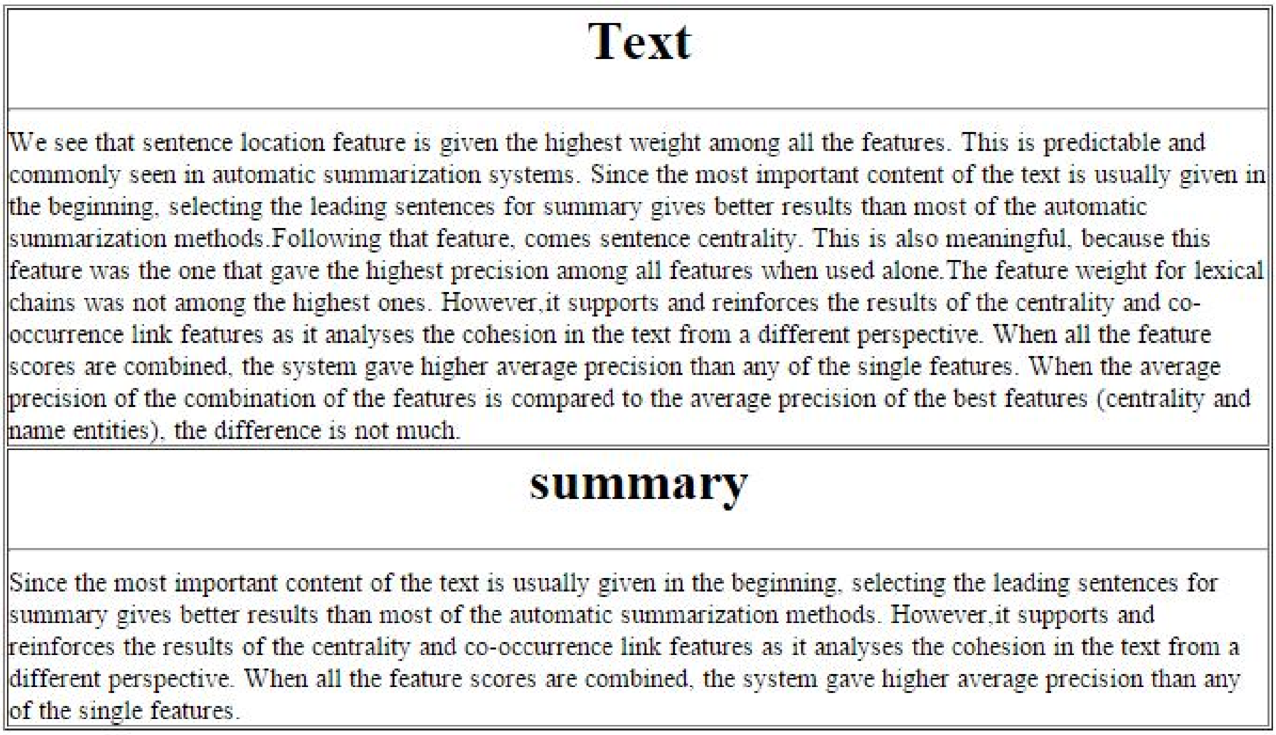
IV.EXPERIMENTAL RESULTS

Here, 40 English manually summarized documents are taken for training purpose which have compression ratio of 30%. Features of these documents are extracted, some weights also included according to GA equation. The best weights (Wi) are used for generating the summary of new 20 different documents.For judge the quality of summary between summarized text of ATS and the manual summary. We measure the system performance in terms of precision from the following formula:

$$P = \frac{S \cap T}{S}$$

Where, P is the precision, T is the manual summary and S is the summary generated by ATS.

There is a screenshot of one small document with its summary. Document and summary is shown in fig. No. 16.



Text

We see that sentence location feature is given the highest weight among all the features. This is predictable and commonly seen in automatic summarization systems. Since the most important content of the text is usually given in the beginning, selecting the leading sentences for summary gives better results than most of the automatic summarization methods.Following that feature, comes sentence centrality. This is also meaningful, because this feature was the one that gave the highest precision among all features when used alone.The feature weight for lexical chains was not among the highest ones. However,it supports and reinforces the results of the centrality and co-occurrence link features as it analyses the cohesion in the text from a different perspective. When all the feature scores are combined, the system gave higher average precision than any of the single features. When the average precision of the combination of the features is compared to the average precision of the best features (centrality and name entities), the difference is not much.

summary

Since the most important content of the text is usually given in the beginning, selecting the leading sentences for summary gives better results than most of the automatic summarization methods. However,it supports and reinforces the results of the centrality and co-occurrence link features as it analyses the cohesion in the text from a different perspective. When all the feature scores are combined, the system gave higher average precision than any of the single features.

Fig .16 Summary of aDocument

V. CONCLUSION AND FUTURE WORK

In this paper, genetic algorithm (GA) is used for automatic text summarization task. Here, these approaches are apply on a sample of some English articles. These approaches have been used the feature extraction criteria. Some of these Features can be used on some other languages for text summarization like presence of brackets, presence of commas, sentence length, sentence position etc. and some text features are language dependent like positive and negative keywords while some other features are language independent.

In the future, ATS may further be extended to multi document summarization. Quality of the summary may further be improved by implementing optimization techniques, some semantic features and some others features given below:-
1. Text summarization tools can be linked with various applications available online for summarizing data.
2. Abstractive approaches to text summarization can be added to improve quality of the summary to large extent.
3. Summarization approaches can be implemented for other multimedia such as audio, video etc.
4. Text summarization task can be extended to otherlanguages also.

REFERENCES.

1. Inderjeet Mani. Automatic summarization, volume 3. John Benjamins Publishing, 2001.
2. Inderjeet Mani and Mark T Maybury. Advances in automatic text summarization. the MIT Press, 1999.
3. P. B. Baxendale, “Machine-made Index for Technical Literature -An Experiment,” Journal of Research and IBM Development, vol. 2, no. 4, pp. 354- 361, October 1958. DilipKumar S. M. and Vijaya Kumar B. P. ,’Energy-Aware Multicast Routing in MANETs: A Genetic Algorithm Approach’, *International Journal of*Computer *Science and Information Security* (IJCSIS), Vol. 2,2009.
4. Dragomir R Radev, Hongyan Jing and et al, “Centroid-based summarization of multiple documents,” Information Processing & Management, Elsevier, vol.40, no.6, pp.919-938, 2004. D.Shama and A.kush,’GPS Enabled EEnergy Efficient Routing for Manet’, International Journal of Computer Networks (IJCN),Vol.3, Issue 3, pp. 159-166,2011.
5. Harold P Edmundson. New methods in automatic extracting. Journal of the ACM (JACM), 16(2):264–285, 1969.

6 Russell, S. J., &Norvig, P. (1995). Artificial intelligence: a modern approach. Englewood Cliffs, NJ: Prentice-Hall International Inc.

7 Yeh, J., Ke, H., Yang, W., &Meng. I. (2005). Text summarization using a trainable summarizer and latent semantic analysis. Information Processing & Management, 41(1), 75-95.

8 Automatic Text Summarization Mohamed Abdel Fattah, and Fuji Ren.

9 Gurmeet Singh and Karun Verma, “A Novel Features Based Automated Gurmukhi Text Summarization System,” International Conference on Advance in Computing, Communication and Information Science, Elsevier, 2014.