# Extractive Text Summarization – An effective approach to extract information from Text

Conference Paper · December 2019

DOI: 10.1109/IC3I46837.2019.9055636

Asha Mishra
G L Bajaj Institute of Technology and Management

# Extractive Text Summarization – An effective approach to extract information from Text

*Asha Rani Mishra*
*Research Scholar, Department of Computer Science, Al Falah University,*
*Faridabad, India*
*asha1.mishra@gmail.com*

*V.K Panchal*
*Prof., Department of Computer Science, Al Falah University,*
*Faridabad, India*
*vkpans@gmail.com*

*Pawan Kumar*
*Prof., Department of Computer Science, BSAITM,*
*Faridabad, India*
*pawan.bhadana79@gmail.com*

***Abstract*-Everyday large volume of data is gathered from different sources and are stored since they contain valuable piece of information. The storage of data must be done in efficient manner since it leads in difficulty during retrieval. Text data are available in the form of large documents. Understanding large text documents and extracting meaningful information out of it is time-consuming tasks. To overcome these challenges, information in the form of text are summarized in with an objective to get relevant knowledge with the help of text mining tools. Summarized text will have reduced size as compared to original one. In this paper, we have tried to highlight major techniques for extracting important information from a given text with the help of topic modeling, key phrase extraction and summary generation .For topic modelling LSI and NMF method is used ,weighted TF-IDF method is used for key phrase extraction while text summary is generated by using LSA and Text Rank method.**

***Keywords—Text Summarization, Extractive methods, Machine Learning, LDA, Text Rank, Ranking, Topic modelling, similarity measures.***

## I. Introduction

Information in the form of unstructured text is available to us in large volume from various sources. While working with text, topic modelling, key phrase extraction and text summarization are the methods that can be used for extracting information from the text. Topic modeling used probabilistic statistical approach to find latent variables which explores latent semantic structures present in text like SVD (Singular Value Decomposition) and LDA (Latent Dirichlet Allocation. NMF method has advantages over SVD in terms of nonnegativity and sparseness [1]. Key phrase extraction describes whole text in few words which gives its overall view. Since there is a urgent need of finding a way which extracts important information in in an effective way in less time. Text summarization provides a way to extract information in a short period of time by reducing the size of the text without changing the original meaning of the original text. It is widely used in text analytics in various domain.

## II. Related Work on Extractive Text Summarization

Automatic text summarization is the technique of producing only relevant, correct and precise summaries from a larger size document by retaining only relevant sentences [2]. Several methods of extraction strategies have been successfully used in the past to retrieve relevant contents as a document summary. Summarizing text can be based on number of input document (Single vs Multiple), content type of original text (Generic vs Query-based), based on category of input text (Genre specific vs Domain-specific), based on purpose (Inductive vs Informative or critical), based on level of linguistic space (Shallow vs Deep) and based on output type (Extractive or Abstractive)[3].Table I shows broad methods of text summarization based on some parameters and Table II shows previous commonly used methods of extractive text summarization.

Table I . Previous research work on the basis of different parameters of Text Summarization

| Researchers | Type of Approach Used | Merits/Demerits of the approach used |
|---|---|---|
| Nenkova et al., 2011 | Generic | **Merits:** Domain knowledge is not considered while generating summary. Tries to visualize homogeneity in a document.<br>**Demerits:** Generic summaries do not have any view of the topic and consider the document as a unique text, so all the information have the same level of importance. |
| Radev et al., 1998; Verma et al., 2007; Wu et al., 2003 | Domain Specific | **Merits:** Domain specific knowledge plays a key role to find sentence selection process.<br>**Demerits:** Difficult to encode domain-specific knowledge. |
| Nenkova et al, 2011 | Query Specific | **Merits:** Information in the generated summary are query specific.<br>**Demerits**: Does not provide exact view of the document's concepts since they focus on the user's query. |
| Aliguliyev, 2009; Ko et al., 2008 | Extractive | **Merits:** Extractive summaries contain important sentences selected from the document without any alteration.<br>**Demerits:** It comes with a risk of producing an inconsistent text since the selected sentences may not share a semantic relation with one another i.e. incoherent summary is produced. |
| Singhal et al., 2010 | Abstractive | **Merits:** Semantic analysis which considers only coherent sentences of the document for combination. It excludes irrelevant sentences from the summary.<br>**Demerits:** Lack of coherency in longer summaries, incorrect reproduction of facts, and lack of novelty in generated sequences, dependent on deep linguistic skills. |

978-1-7281-5529-6/19/$31.00 ©2019 IEEE

| | | |
|---|---|---|
| | | |
| S. Gholamrezazadeh, et al. 2009 | Indicative | Merits: Present the main idea of the entire document, it gives the user a quick view from the original text.<br>Demerits: So, it may not contain all important factual content. |
| C.T. Shubhangi et al.,2014 | Informative | **Merits:** Express the important concise information of the original text to the user.<br>**Demerits**: Difficult to find informative sentences for summary. |
| Saggion et al ., 2013 | Multiple document | Merits: Topics are redundant in multiple document<br>Demerits: Redundancy, sentence ordering, temporal dimension and co-reference are major issues which needs to be addressed. |
| Svore et al. ,2007 | Single document | **Merits:** Systems produced a summary from a single source document.<br>**Demerits:** Type of text does not have an impact as they do not consider domain specific knowledge. |

TABLE II. PREVIOUS RESEARCH BASED ON VARIOUS EXTRACTION STRATEGIES FOR TEXT SUMMARIZATION

| **Researchers** | **Type of approach** | **Concept(s) used** |
|---|---|---|
| Nenkova et al.,2005,<br>Filatova et al., 2004,<br>Fung et.al, 2006,<br>Galley, 2006,<br>Hovy et al., 1998 | Frequency based | Word probability:<br>Used as measure of importance of a word.<br>Formula: Prob.(word)=Freq.(word)/total words<br>where word € document<br>TF-IDF: Reduces the impact of common occurring words which are more frequent in nature by relating to its proportional frequency in the document set.<br>Formula: TF-IDF = TF * IDF [4]<br>**Merits:** Simple, easy and fast to compute<br>**Demerits:** Does not consider uniqueness of the word. Redundancy of information is extremely high. |
| Binwahlan et al. ,2009,<br>Bossard et al.,2011<br>Suanmali et al. ,2009; 2011<br>Babar et al., 2015<br>Jagadeesh et al., 2005<br>Gunawan et al., 2017 | Feature based | Identifies important features(fi)which will help to find relevant sentences for summary like sentence length, cue method etc. and assigns suitable weight(wi) to it.[5][6]<br>Formula: Sentence Score (SS)=∑(wi) * (fi) for i=1 ,2, 3…...n<br>Example: PyTeaser<br>**Merits:**<br>PSO method is used, Fuzzy rules are designed using the features. Uses Fuzzy reasoning, used both sentence and word level, used labelled sentence level features like title, sentence length, sentence position and keyword frequency.<br>**Demerits:**<br>Does not considers semantics of the sentence. |
| Hannah et al.,2014 | Machine learning based | Suitable for large set of training data consisting of documents along with the summary extract as its label for supervised learning whose objective is to classify a test sentence into predefined classes as summary sentence and non-summary sentence.<br>Popular supervised machine learning algorithms used for text summarization are Classifier based on Naïve Bayes, Decision Tree, Maximum entropy, Neural networks, Support Vector Machine.[7][8][9][10]<br>On the other hand, unsupervised approaches generate summaries without needing of training data. Hidden Markov Model [11], CRF . Clustering and Deep learning techniques (RBM, Autoencoder, Convolutional network, RNN) are instances of unsupervised learning technique.<br>**Merits:** Neural network model is used to generated summaries. Used a decision tree.<br>**Demerits:**<br>Effective in learning important the features that are essential to make a summary but training corpus is different from language to language and is not fixed for every document. Supervised approach is a feature dependent approach i.e. annotations or labelling should be done to the data to be properly trained. |
| Kleinberg, 1999<br>Brin et al., 2012<br>Wan et al. 2006<br>Hariharan et al.2009 | Graph based | Nodes are the sentences and edges define interrelatedness between two sentences. A numeric score is assigned to the edges based on similarity or how much there is relatedness between each sentence. Cosine Similarity, Jaccard coefficient, Euclidean distance are the popular ones. HITS (Hyperlinked Induced Topic Search, Google's PageRank (GPR)<br>**Merits:**<br>Performs good on better for good feature matrix. Assigned different weights to document links.<br>Ranked sentences are non-redundant.<br>**Demerits:** Important issues to be solved are polysemy and the correct interpretation of meaning of phrases.<br>Sentence similarity is used to model graph. |
| Steinberger et al., 2004,<br>Gong et al. 2001<br>Lee et al. 2009 | Algebric based | This method is vector based for extracting a representation of text semantics on the basis of occurrence of words in the text data. Main idea is to map high dimensional data to lower dimensional space without any significant loss of information. For this Topic Modeling is used can be used for dimension reduction can be used in text summarization tasks. Weighting factors both local and global can be used for the value of each entry in word sentence matrix.<br>**Merits:**<br>Clustering of topics for sentence selection. |

| | | |
|---|---|---|
| | | NMF method has advantages over SVD in terms of nonnegativity and sparseness<br>**Demerits:**<br>Inefficient representation, due to its distributional nature. A topic may need more than one sentence. Mostly used with dimension reduction techniques. |

## III. General flow of an Extraction-based (Single document) Summarization

Extractive based method identifies key phrases from the input documents to generate summary. Original sentences are not modified in extractive based method. Main phases are shown in the Figure 1 shown below

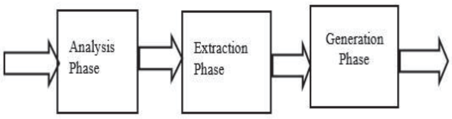

Figure 1: Architecture of Extraction based Methods

**Algorithm for Extractive based Summarization**

Input: Text

*Analysis Phase:*

1. Pre-processing-

*Level 1:* Block level splitting of the text (phrases, sentences, paragraphs)

*Level 2:* Word level splitting of blocks(tokenization)

*Level 3:* Word normalization (Lemmatization, stemming)

*Level 4:* Removal of stop words, POS tagging, Named Entity Recognition, Extraction of terms and keywords

2. Calculation of Similarity measures –To find relatedness between the sentences to produce consistent flow.

3. Weighting of the sentences- Depending upon the feature representation a numeric value is assigned to the sentences.

*Extraction Phase:*

Selection of sentences- Ranking method can be used for this. Higher ranked sentences are preferred for summary.

*Generation Phase-*

Reranking can be done using similarity measures to minimize redundancy.

Output: Summary of the input text

## IV. Results

Following text is taken as input to perform topic modeling, key phrase extraction. Results are shown in the tables III and IV given below.

"Everyday large volume of data is gathered from different sources and are stored since they contain valuable piece of information. The storage of data must be done in efficient manner since it leads in difficulty during retrieval. Text data are available in the form of large documents. Understanding large text documents and extracting meaningful information out of it is time-consuming tasks. To overcome these challenges, text documents are summarized in with an objective to get related information from a large document or a collection of documents. Text mining can be used for this purpose. Summarized text will have reduced size as compare to original one. In this review, we have tried to evaluate and compare different techniques of Text summarization."

Table III. Result of Topic Modelling and Key Phrase extraction on the input text

| | |
|---|---|
| Topic Modeling (without weights) | Topic #1 without weights<br>['since', 'data', 'must', 'retrieval', 'difficulty', 'efficient', 'lead', 'manner', 'storage', 'text', 'contain', 'everyday', 'gather', 'valuable', 'piece', 'volume', 'source', 'store']<br>Topic #2 without weights<br>['text', 'document', 'large', 'data', 'available', 'form', 'information', 'summarize', 'compare', 'meaningful', 'consuming', 'extract', 'understand', 'time', 'task', 'different'] |
| Topic Modeling using LSI | Topic #1 with weights<br>[('different', 0.02), ('data', 0.02), ('since', 0.02), ('technique', 0.02), ('storage', 0.02)]<br>Topic #2 with weights<br>[('form', 0.029999999), ('available', 0.029999999), ('meaningful', 0.029999999), ('task', 0.029999999), ('extract', 0.029999999)] |
| Topic Modeling using NMF | Topic #1 with weights<br>[('document', 0.53), ('large', 0.31), ('text', 0.23), ('available', 0.22), ('form', 0.22), ('information', 0.18), ('data', 0.16), ('consuming', 0.11), ('task', 0.11), ('time', 0.11), ('extract', 0.11), ('understand', 0.11), ('meaningful', 0.11), ('related', 0.1), ('collection', 0.1), ('get', 0.1), ('challenges', 0.1), ('overcome', 0.1), ('objective', 0.1), ('summarize', 0.08)]<br>Topic #2 with weights<br>[('compare', 0.35), ('reduce', 0.22), ('original', 0.22), ('size', 0.22), ('one', 0.22), ('text', 0.2), ('summarize', 0.18), ('review', 0.17), ('evaluate', 0.17), ('summarization', 0.17), ('technique', 0.17), ('try', 0.17), ('different', 0.13)] |

| | |
|---|---|
| Key Phrase extraction using weighted TF-IDF method | [('form',0.5769999999999996), ('large documents', 0.5769999999999996), ('text data', 0.5769999999999996), ('large text documents', 0.5769999999999996), ('meaningful information', 0.5769999999999996), ('time-consuming tasks', 0.5769999999999996), ('different techniques', 0.5769999999999996), ('review', 0.5769999999999996), ('text summarization', 0.5769999999999996), ('different sources', 0.4759999999999998)] |

TABLE IV. RESULT OF TEXT SUMMARIZATION ON THE INPUT TEXT

| **Method Used** | **LSA** | **Text Rank Method** |
|---|---|---|
| Total sentences in the text | 8 | 8 |
| Sentence/Ranking Score | [3.78 3.01 2.11 2.22 4.9 0.44 1.52 1.97] | [(0.14590720885501199, 2), 0.13703086865833658, 4), (0.13077170594965404, 3), (0.12514209311536181, 0), (0.11925738491595521, 6), (0.11887192911445364, 7), (0.11228742359741933, 5), (0.11073138579380756, 1)] |
| Index of the top 3 sentences selected for the summary | [014] | [2, 3, 4] |

## V. CONCLUSION AND FUTURE WORK

This paper tries to highlight the related research work done in the past years for text summarization using extractive methods. Commonly used extractive approaches are discussed along with their merits and demerits. They are simple as compare to abstractive summaries but are less generalized as compared to abstractive methods since they did not exploit more linguistic and semantic understanding, inference and natural language generation while generating summary. So, to generate a summary carrying more concise information is the main challenge in these methods. More effort should be done in these methods to generate summary having right informative content, proper flow, less redundancy, coherency between the sentences, and appropriate size. Graph based (Text Rank method) can generate query specific and domain specific summarization very well but finding proper relatedness function is a key issue. Latent semantic Analysis (LSA) method captures semantic relation to generate summary but cannot handle problems like polysemy etc. Finding proper evaluation metrices to analyze the performance is also a challenging task. With the recent advancement in the field of machine learning, deep learning and natural language processing and other domains of AI efficient summary can be generated for larger volume of data along with the proper feature representation and better semantic understanding. More research can be done for proper sequencing of the generated summary that involves that include interactions between sentences with the help of deep learning techniques.

## VI. REFERENCES

[1] J. H. Lee, S. Park, C. M. Ahn, D. Kim, "Automatic generic document summarization based on non-negative matrix factorization" Journal on Information Processing and Management, vol. 45(1), 2009, pp. 20–34.

[2] . E. Lloret and M. Palomar, "Text summarisation in progress: a literature review," in Springer, no. April 2011, pp. 1–41, Springer, 2012.

[3] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive techniques," Journal of Emerging Technologies in Web Intelligence, vol. 2, no. 3, pp. 258–268, 2010.

[4] R. Ferreira, L. De Souza Cabral, R. D. Lins, G. Pereira E Silva, F. Freitas, G. D. C. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, "Assessing sentence scoring techniques for extractive text summarization," Expert Systems with Applications, vol. 40, no. 14, pp. 5755–5764, 2013.

[5] G.Forman., "An extensive empirical study of feature selection metrics for text classification". Journal of machine learning research, 2003. 3(Mar): p. 1289-1305.

[6] B. Trstenja, S. Mika, D.Donko, "KNN with TF-IDF Based Framework for Text Categorization",, 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013

[7] Gholamrezazadeh, S., Salehi, M.A., Gholamzadeh, B." A comprehensive survey on text summarization systems. "2nd International Conference on Computer Science and its Applications, Jeju, Korea (South), pp. 1–6 (2009).

[8] M. Osborne, "Using maximum entropy for sentence extraction," In Proceedings of the AssociaCL-02 Workshop on Automatic Summarization, ACL, 2002, pp. 1–8.

[9] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou. "Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization". In AAAI, pages 2153–2159, 2015.

[10] Z. Cao, F. Wei, S. Li, W. Li, M. Zhou, and H. Wang. "Learning Summary Prior Representation for Extractive Summarization." In ACL (2), pages 829–833, 2015

[11] John M,Dianne P, "Text summarization via hidden Markov models," Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval,2001