# Survey of Automatic Text Summarization Techniques & Algorithms

**K. Gowri[1] Dr. R. Manickachezian[2]**
[1]Research Scholar [2]Associate Professor
[1,2]Department of Computer Science
[1,2]N.G.M College, India

*Abstract*— Today in the textual information is quickly growing and is out there in many alternative domains. The knowledge is out there in profusion for each topic on worldwide internet, compression the necessary data within the sort of outline would profit a distinct application. Hence, there's growing interest among the analysis community for developing new approaches to mechanically summarize the text. Automatic text report could be a technique that compresses giant text to the most words text which incorporates the necessary data. During this paper to associate in Nursingalysis totally different algorithms are developed for an automatic outline generation that implements variety of machines learning and improvement techniques, a survey of recent text report extractive approaches developed within the recent analysis techniques. What is more, analysis results of extractive report approaches are conferred on some shared Reuters datasets. Finally, this paper concludes with the discussion of helpful future directions which will facilitate analysers to spot areas wherever any research is required.

***Key words:*** Text Summarization Techniques

## I. INTRODUCTION

Text summarization is the problem of creating a short, accurate, and fluent summary of a longer text document. Automatic text summarization methods are greatly needed to address the ever-growing amount of text data available online to both better help discover relevant information and to consume relevant information faster. Automatic summarization of text works by first calculating the word frequencies for the entire text document. Then, the 100 most common words are stored and arranged. Each sentence is then scored based on how many high frequency words it contains, with higher frequency words being worth more. Finally, the top X sentences are then taken, and sorted based on their position in the original text.

Business leaders, analysts, paralegals, and tutorial researchers ought to comb through immense numbers of documents on a daily basis to stay ahead, and an oversized portion of their time is spent simply working out what document has relevancy and what isn't. By extracting vital sentences and making comprehensive summaries, it's attainable to quickly assess whether or not or not a document is price reading. Automatic text account is additionally helpful for college kids and authors. Imagine having the ability to mechanically generate Associate in nursing abstract based mostly for your analysis paper or chapter in an exceeding book in an exceedingly clear and aphoristic means that's trustworthy to the first supply material.

Many document report strategies area unit supported standard term weight approach for choosing the valid sentences. a collection of frequencies and term weights supported the quantity of occurrences of the words is calculated. Report strategies supported linguistics analysis conjointly use term weights for final sentence choice. The term weights usually used don't seem to be directly derived supported any mathematical model of term distribution or connectedness. Mathematically characterize the connectedness of terms in an exceedingly document. This model is then wont to extract necessary sentences from the documents.

Fig 1 show the text summarization over flow diagram, first load the dataset (corpus) than pre-process the document based on stop word or steaming. The feature selection techniques used to weight each term based on frequency, finally apply the text classification algorithm to get the result.

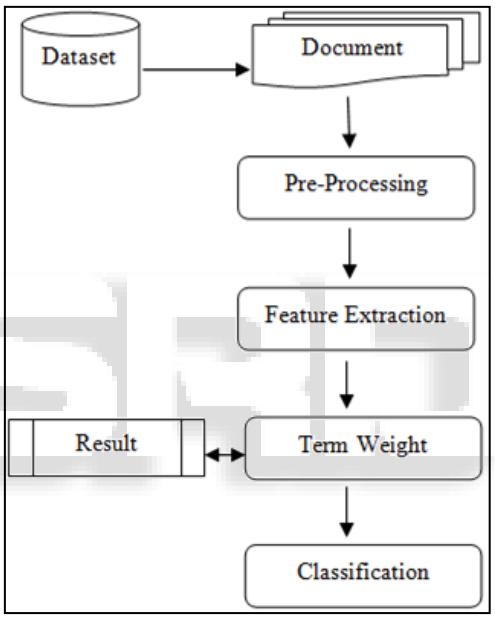


Fig. 1: Overview of Text Summarization

Summary can be defined as a brief and accurate way of representing the important concepts of the given source documents. Humans, during the process of text summarization, understand the concept of source document and create a summary which conveys the essence of the document whereas in automated systems this is a complex task. As the quantity of information available in electronic format continues to grow, research into automatic text summarization has taken huge importance. There are two types of summary Extractive and Abstractive. Abstractive summary represents use of. (NLP) whereas Extractive summary is based on copying exact sentences from source document. Presently it is not possible that the computer can understand every aspect behind Natural Language processing. So, our Scope is limited to Extractive based summary.

Automatic summarization aims at manufacturing a taciturn, condensed illustration of the key data contained in associate degree data supply for a specific user and task.

In addition to developing higher theoretical foundations and improved characterization of summarization

 

issues, any work on correct analysis ways and summarization resources, particularly corpora, is of nice interest. Generally, text summarization methods are classified broadly into two categories. One class relies on victimization applied mathematics live to derive a term-weighting formula. The other relies on victimization linguistics analysis to spot lexical cohesion within the sentences. This approach isn't capable of handling massive corpora. Both the approaches finally extract the vital sentences from the document assortment. Sentences from the document collection.

## II. RELATED WORK

The text summarization and retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the various offered documents are relevant to a given individual or user. Without knowing what might be within the documents, it's troublesome to formulate effective queries for analyzing and extracting helpful data from the info. More and additional courts round the world area unit providing on-line access to judgments of cases, each past and gift. With this exponential growth of on-line access to legal judgments, it's become progressively vital to supply improved mechanisms to extract data quickly and gift rudimentary structured information instead of mere information to the legal community. Automatic text summarization makes an attempt to handle this downside by extracting data content, and presenting the foremost vital content to the legal user.

### *A. Single Document Summarization*

Automatic summarizers typically identify the most important sentences from an input document. Major approaches for determining the salient sentences in the text are term weighting approach [1], symbolic techniques based on discourse structure [2], semantic relations between words [3] and other specialized methods [4].While most of the account efforts have targeted on single documents, many initials comes have shown promise within the account of multiple documents. The construct of multi-document, multilingual and cross-language information retrieval. All these methods are tried singly as well as in combinations. From the higher than studies, we understand that the automatic extraction systems need more sophisticated representations than single words. The best combination is chosen on the idea of the best average proportion of sentences common within the automatic extracts and also the target extracts.

### *B. Automatic Extraction of Sentence*

Automatic summarizing via sentence extraction operates by locating the most effective content bearing sentences during a text. Extraction of sentences can be simple and fast. The drawback is that the ensuing passage may not be clear. It sacrifices the coherence of the supply for speed and practicable. Hence, we need to use appropriate ways to undertake this downside and gift the outline during a lot of easy manners. The assumption behind extraction is that there is a set of sentences, which present all the key ideas of the text, or at least a majority of these ideas. The goal is initial to spot what very influences the importance of a sentence, what makes it important. The next step is to extract important sentences based on the syntactic, semantic and discourse analysis of the text. Systems designed on a restricted domain show promising results.

### *C. Extraction-Based Summarization*

The techniques for automatic extraction can be classified into two basic approaches [5]. The first approach is based on a set of rules to select the important sentences, and the second approach is based on a statistical analysis to extract the sentences with higher weight.

#### *1) Rule-Based Approach*

This method uses the facts that determine the importance of sentence as encoded rules. The sentences that satisfy these rules are the ones to be extracted.

#### *2) Statistical Approach*

In distinction to the manual rules, the statistical approach basically tries to automatically learn the rules, which predict a summary-worthy sentence. Statistics-based systems area unit empirical, re-trainable systems, which minimize human effort. Their goal is to identify the units in a sentence which influence its importance and to learn the dependency between the occurrence of units and the significance of a sentence. In this framework, every sentence is allotted a score that represents the degree of appropriateness for inclusion in an exceeding outline.

### *D. System for Automatic Extraction*

The following are the factors to be considered in the process of automatic extraction of sentences from a document collection.

#### *1) Length of an Extract*

The [6] postulate that about 20% of the sentences in a text could convey all the basic ideas about it. Since abstracts are much shorter than this proportion, the length of extracts should lie between the length of an abstract and the Morris's figure.

#### *2) Proportion*

The predefined percentage (usually 10%) of the number of sentences of the document should be selected. This technique is nice for ordinarily sized documents however can manufacture long extracts for long documents.

#### *3) Oracle Method*

If a target extract is on the market, select the same number of sentences. In addition, it is intuitive that a computer extract will need more sentences than the perfect extract in order to have a good point of coverage and coherence. An advantage of the oracle technique is that the systems are often "trained" from the target extracts so the optimum varieties of sentences are often foretold from the test documents.

#### *4) Fixed Variety of Sentences*

Here the length of Associate in Nursing extract is usually constant (typically, 10-15 sentences) regardless of the size of the documents. This technique is closer to human-produced abstracts. It favors shortness, but the problems in the previous methods continue.

#### *5) Sentences above a Certain Threshold*

For a sentence to be included in the extract, it suffices to have a score which is reasonable enough. This is one way of trade-off between the extremes of the previous methods, but it requires determination of a threshold.

 

*6) Mathematical Formula*

The number of extracted sentences is an increasing function of the number of sentences in the text, but it does not grow linearly. Hence, comparatively few sentence area unit further once the text is huge, and fewer still for a much bigger one. This is probably one of the best methods as it prevents a size explosion. It caters to huge documents as well.

*7) Length of a Sentence*

It's going to be explicit that sentences that area unit too short or too long area unit typically not ideal for Associate in Nursing abstract, and therefore for an extract as well. This is typically spoken [31] as sentence cut-off feature. It penalizes short (less than 5-6 words) and long sentences either by reducing their score, or by excluding them fully.

## III. TEXT SUMMARIZATION USING MACHINE LEARNING TECHNIQUES

Extractive summarizers [7][8] aim at picking out the most relevant sentences in the document while also maintaining a low redundancy in the summary.

### *A. Cluster based Methods*

Cluster based methods measures relevance or similarity between each sentence in a document with that of sentences selected for summary. Summaries address onto different "themes" appearing in the documents, which is incorporated through clustering. Clustering based methods become essential to generate a meaningful summary. Documents square measure sometimes written specified they address totally different topics one when the opposite in associate degree organized manner. They are commonly broken up explicitly or implicitly into sections. This organization applies even to summaries of documents. It is intuitive to assume that summaries ought to address totally different "themes" showing within the documents. Some summarizers incorporate this aspect through clustering. If the document assortment that outline is being created is of all totally different topics, document clump becomes nearly essential to get a purposeful outline

### *B. Graph theoretic Approach*

Representation is an extractive summarization model, which provides a method to identify themes in the document. Preprocessing steps, namely, stop word removal and stemming are done before, to obtain graphical view of the documents. Sentences in the documents form nodes of an undirected graph. An edge is drawn among the nodes if sentences share some common words, or whose (cosine, or such) similarity is above some threshold. Graphic representation yields partitions indicating distinct topics covered in the documents. Identification of nodes with high cardinality forms higher preferred sentences to be included in the summary.

### *C. LSA Method*

Singular Value Decomposition (SVD) [9] is a very powerful mathematical tool that can find principal orthogonal dimensions of multidimensional data. It has applications in several areas and is thought by completely different names: Karhunen-Loeve remodel in image process, Principal Component Analysis (PCA) in signal processes and Latent Semantic Analysis (LSA)in text processing. It gets this name LSA because SVD applied to document word matrices, groups documents that are semantically related to each other, even when they do not share common words.

Words that sometimes occur in connected contexts are connected within the same singular house. This technique may be applied to extract the topic-words and content-sentences from documents. The advantage of mistreatment LSA vectors for summarization instead of the word vectors is that abstract (or semantic) relations as diagrammatical in human brain area unit mechanically captured within the LSA, while using word vectors without the LSA transformation requires design of explicit methods to derive conceptual relations. Since SVD finds principal and reciprocally orthogonal dimensions of the sentence vectors, selecting out a representative sentence from every of the scale ensures connectedness to the document, and orthogonality ensures non-redundancy. It is to be noted that this property applies solely to information that has principal dimensions inherently but, LSA would probably work since most of the text data has such principal dimensions thanks to the range of topics it addresses.

### *D. Preprocessing*

The preprocessing steps commonly performed in automatic summarization are word stemming, stop words removal, text segmentation and query expansion. These routines are also quite common for other NLP tasks such as document retrieval, information extraction and machine translation.

*1) Stemming*

Stemming is the process of reducing the inflected forms of a word to a root form. It is commonly used in information retrieval tasks to reveal semantic similarity between different morphological variants of a word. For example, verbs "play", "plays" and "played" all have the same root form "play" and thus might be represented by only one feature (i.e. term). This makes it possible to detect similarities between contexts that contain these words. In addition, word stemming reduces the overall number of features, making the data less sparse.

*2) Stop Words Removal*

Stop words are high-frequency words of a language that don't carry any particular information on their own. Such words are removed at the preprocessing section to cut back the amount of options. Closed category words like pro-nouns, articles, prepositions and conjunctions are usually enclosed in stop words lists. In addition, some of the frequently used open class words such as auxiliary verbs are also included. Stop word lists of various granularities are freely available on the Web and often utilized in summarization. During the removal procedure all the words that appear in a list of stop words are removed from the source documents.

*3) Query Expansion*

In query-oriented summarization the content of a query provides valuable clues for deciding which parts of the documents are important. However, queries are often very short and it might be beneficial to expand them in order to obtain more clues. Several methods using a variety of external sources have been proposed for this task. The general idea behind many of them is to use external resources such as for example Word Net for identifying terms that are related to the terms in a query.

 

*4) Text Segmentation*

Text segmentation is essential for extraction-based summarization. Segmentation divides an input document into separate textual contexts. Ideally, a context should express a relatively independent and standalone piece of information. In extraction-based summarization, sentences are considered as contexts. Although it may seem that finding sentence boundaries is a trivial task, in fact, punctuation ambiguities make it rather hard. Many natural languages use periods to indicate sentence boundaries but periods can also signal abbreviations, initials or ordinal numbers. Extractive summarization requires high accuracy sentence boundary detection methods because even a single incorrectly separated incomplete sentence will greatly reduce the coherence and readability of a summary. Several approaches have been applied for this problem. Majority of the approaches are either rule-based or rely on machine learning to identify sentence boundaries.

*5) Feature Selection*

Different words and phrases carry different amount of information. A variety of methods are applied to discard unimportant features and to weight the remaining ones in correspondence with their importance. The basic method mentioned previously in the report is stop words removal, which utilizes a predefined list of unimportant lexical features. More sophisticated weighting techniques such as TF-IDF and log-likelihood ratio use statistical information about distribution of words in a set of documents.

The TF-IDF weight is the product of two components; the term frequency TF and the inverse document frequency IDF. The TF determines the importance of a word for a given document, while the IDF indicates the importance of a word over the whole set of documents. A word that occurs often in a specific document but rarely in other documents is considered to be relevant for this document and, consequently, receives a high weight value.

*6) Similarity Measures*

Textual similarity is a complex concept that can be defined as the semantic relatedness of two textual contexts. Two contexts are considered similar if they focus on the same or related concepts, actors or actions. Measuring the similarity between textual contexts is an intermediate step for many NLP applications. Most of the research regarding text similarity has been done by IR community, where assignment of similarity between a query and a document is a core task. In automatic summarization, similarity metrics are used for centrality-based context selection and in identification of redundant contexts. In general, similarity measures are either corpus-based or knowledge-based. Both of them have been used in extractive summarization. Corpus-based measures use term frequencies observed in a corpus to relate contexts to each other, while knowledge-based methods utilize predefined semantic relations between terms obtained from lexical resources.

*7) Content Selection*

Selection of sentences is the core process in extractive summarization. The goal of the choice procedure is to spot a collection of sentences that contain vital data. Three criteria are optimized when selecting the sentences: relevance, redundancy and length. Relevance determines the importance of the information contained in a summary with respect to the topics covered in the source documents or a query in case of query-focused summarization. Redundancy measures the information overlap between the sentences selected for the summary. Given a restricted summary length, summarization systems try to maximize the relevance while minimizing the redundancy. The task of content selection is to identify which sentences in the source documents are worth taking into a summary. The problem can be handled either in a supervised or unsupervised manner.

1) Supervised Methods

Supervised techniques use a classifier trained on a set of documents coupled with corresponding extracts. Extracts make it possible to label sentences in the documents with a binary value: 1 - a sentence is included in the extract, 0 - a sentence is not included in the extract. In addition, each sentence should be represented by a feature vector. A feature vector is constructed using a predefined set of features that should provide enough information to determine whether a particular sentence is worth including in a summary.

2) Unsupervised Methods

The unsupervised content selection methods identify salient sentences without training on a labeled set of documents. Most of the unsupervised summarization algorithms are either centroid-based or centrality-based. The general idea behind centroid-based algorithms is to select sentences that contain informative words, also referred to as topic signatures. Stop words like articles and pronouns are usually ignored. The in formativeness of the remaining words is calculated using popular weighting schemes such as TF-IDF or log-likelihood ratio.

*8) Redundancy Removal*

Redundancy is a major issue in multi-document summarization where several documents on the same topic may have a substantial information overlap. Then, the selection of the most relevant sentences will yield a set of sentences with redundant information. Extract that consists of relevant but very similar sentences is not good. The joint optimization of both relevancy and redundancy is a complex task because properties of individual sentences are dependent on other sentences included in the summary. Some of the earlier multi-document summarization approaches handle these optimizations separately. For example, clustering is used to obtain groups of similar sentences and then the most authoritative contexts from each group are selected [10]. This way the clustering step minimizes redundancy and the selection of authoritative contexts from each cluster maximizes relevance.

## IV. DATASET AND PERFORMANCE METRICS

Even though these numbers are not comparable to other results since a subset and not the complete Reuters 21578 split was used, they provide still interesting Insights. Especially the fact, that for the same weighting function and the same dimensionality, it happens that, e.g., the breakeven value is higher compared to another function but the eleven-point precision is lower, compared to the same function. It also shows that" MSF" could be an interesting alternative to chi-square and information gain, not only for feature selection in text classification, but also to weight the importance of features in other classification tasks.

 

*1) Cross-Validation*
Cross -validation may be a technique for estimating the generalization performance of a prophetic model. The main idea behind CV is to split data, once or several times, for estimating the risk of each algorithm: Part of data (the training sample) is used for training every algorithmic rule, and the remaining part (the validation sample) is used for estimating the risk of the algorithm. Then, CV selects the algorithmic rule with the tiniest calculable risk. Cross validation is an alternative to random sub sampling.

*2) Confusion Matrix*
A binary classification model classifies each instance into one of two classes; say a t rue and a false class. This gives rise to four doable classifications for every instance: a real positive, a true negative, a false positive, or a false negative. This situation may be represented as a confusion matrix. The confusion matrix juxtaposes the ascertained classifications for a development (columns) with the anticipated classifications of model (rows) .The classifications that lie on the most important diagonal of the table area unit the right classifications, that is, the true positives and the true negatives. The other fields signify model errors. For a perfect model would only see the true positive and true negative fields filled out, the other fields would be set to zero. It is common to decision true positives hits, true negatives correct rejections, false positive.

*3) Receiver Operating Curves (ROC)*
Central to constructing, deploying, and using classification models is the question of model performance assessment. Traditionally this can be accomplished by mistreatment metrics derived from the confusion matrix or contingency table. However, it's been recognized that (a) a scalar may be a poor outline for the performance of a model especially once deploying non-parametric models like artificial neural networks or call trees and (b) some performance metrics derived from the confusion matrix area unit sensitive to knowledge anomalies like category skew. Recently it's been ascertained that Receiver operational Characteristic (ROC) curves visually convey an equivalent info because the confusion matrix during a far more intuitive and strong fashion. ROC curves area unit two-dimensional graphs that visually depict the performance and performance trade-off of a classification model. ROC curves were originally designed as tools in discipline to visually verify optimum operational points for signal discriminators. Two new performance metrics ought to be introduced here so as to construct mythical monster curves (they are outlined here in terms of the confusion matrix), the true positive rate (TPR) and the false positive rate (FPR).

*4) Precision*
Precision is that the variety of True Positives divided by the quantity of True Positives and False Positives. Put otherwise, it is the number of positive predictions divided by the total number of positive class values predicted. It is additionally referred to as the Positive prophetic worth (PPV).

*5) Recall*
Recall is that the variety of True Positives divided by {the variety the amount the quantity} of True Positives and also the number of False Negatives. Put otherwise it's the quantity of positive predictions divided by the quantity of positive category values within the take a look at knowledge. It is additionally referred to as Sensitivity or actuality Positive Rate.

*6) F1 Score*
The F1 Score is the 2*((precision*recall) / (precision + recall)).It is additionally referred to as the F Score or the F measure. Put otherwise, the F1 score conveys the balance between the precision and also the recall.

## V. CONCLUSION

This paper emphasized various extractive approaches for single and multi-document summarization. We represented a number of the foremost extensively used strategies like topic illustration approaches, frequency-driven methods, graph-based and machine learning techniques. Although it is not feasible to explain all diverse algorithms and approaches comprehensively in this paper, we think it provides a good insight into recent trends and progresses in automatic summarization methods and describes the progressive during this analysis area. These three recent developments in summarization complement traditional topics in the field that concern intermediate representations and the application of appropriate machine learning methods for summarization

## REFERENCES

[1] J. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, vol. 24, no.5, pp.513-323, 1988.
[2] D. Marcu, "From Discourse Structures to Text Summaries", Proc. of the ACL 97/EACL-97 Workshop on intelligent scalable Text Summarization, pp.82-88, Madrid, Spain, 1997.
[3] R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization", Proc. of the ACL Workshop on Intelligent Scalable Text summarization, pp. 10-17, Madrid, Spain, 1997.
[4] D.R. Radev, H. Jing, and M. Budzikowska, "Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies", Proc. of ANLP-NAACL Workshop on Summarization, pp. 21-30, Seattle, Washington, April, 2000.
[5] B. Georgantopoulos, Automatic Summarizing Based on Sentence Extraction: A Statistical Approach, MSc in Speech and Language Processing Dissertation, University of Edinburgh, 1996.
[6] D. Radev, E. Hovy, K. McKeown, Introduction to the Special Issue on Summarization, Computational Linguistics, vol. 28, no. 4, Association for Computing Machinery, 2002.
[7] Madhavi K. Ganapathiraju, "Overview of summarization methods", 11-742: Self-paced lab in Information Retrieval, November 26, 2002.
[8] Klaus Zechner, "A Literature Survey on Information Extraction and Text Summarization", Computational Linguistics Program, Carnegie Mellon University, April 14, 1997.
[9] Madhavi K. Ganapathiraju, "Overview of summarization methods", 11-742: Self-paced lab in Information Retrieval, November 26, 2002.
[10] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin, "Towards multidocument

 

summarization by reformulation: Progress and prospects," in The National Conference On Artificial Intelligence, vol. pages. John Wiley Sons Ltd, 1999, pp. 453–460.



 