Document Summarization: A Survey

Naveen P Pandurangi
Computer Science and Engineering
Vivekananda Institute OF Technology
Bangalore, India
naveenpandurangi@yahoo.com

Srivatsa V Jamadagni
Computer Science and Engineering
Vivekananda Institute OF Technology
Bangalore, India
Srivatsa.jamadagni@gmail.com

Nikhil Chandran
Computer Science and Engineering
Vivekananda Institute OF Technology
Bangalore, India
nikhilchandren4937@gmail.com

Mohammed Abu Talha Ahmed
Computer Science and Engineering
Vivekananda Institute OF Technology
Bangalore, India
Mohammedshahid870@gmail.com

Abstract— **Document Summarization is the technique by which the huge parts of content are retrieved. The Document Summarization plays out the summarization task by unsupervised learning system. The significance of a sentence in info content is assessed by the assistance of Simplified Lesk calculation. As an online semantic lexicon WordNet is utilized. Word Sense Disambiguation (WSD) is a critical and testing system in the territory of characteristic dialect handling (NLP). A specific word may have distinctive significance in various setting. So, the principle task of word sense disambiguation is to decide the right feeling of a word utilized as a part of a specific setting. To begin with, Document Summarization assesses the weights of the considerable number of sentences of a content independently utilizing the Simplified Lesk calculation and orchestrates them in diminishing request as indicated by their weights. Next, as indicated by the given level of rundown, a specific number of sentences are chosen from that requested rundown.**

Keywords- Document Summarization; Natural Language Processing; Lesk Algorithm; Word Net ;NLTK;

1. INTRODUCTION

Document Summarization, is the plan to get an important data from a huge amount of information. The amount of data accessible on internet is increasing every day so it turns space and time expanding matter to deal with such huge amount of information. So, managing that large amount of data is makes a major problem in different and real data taking care of uses. The Document Summarization undertaking makes the users simpler for various Natural Language applications, like, Data Recovery, Question Answering or content decreasing etc. Document Summarization assumes an inescapable part by creating significant and particular data from a lot of information.Filtering from heaps of reports can be troublesome and tedious. Without a summary or rundown, it can take minutes just to make sense of what the people will discuss in a paper or report. So, the Document Summarization that concentrates a sentence from a content record, figures out which are the most imperative, and returns them in a readable and organized way. Document Summarization is a piece of the field natural language processing, which is the manner by which the PCs can break down, and get importance from human dialect.

Document Summarization that uses the classifier structure and its rundown modules to look over huge amount of reports and returns the sentences that are helpful for producing a summary. Programmed outline of content works by taking the overlapping sentences and synonymous or sense from wordnet most overlapping sentences are considered as high score words. The higher recurrence words are considering most worth. And the top most worth words and are taking from the content and sorted according to its recurrence and generate a summary.

PHASES OF DOCUMENT SUMMARIZATION

There are basic three phases of document summarization as follows:

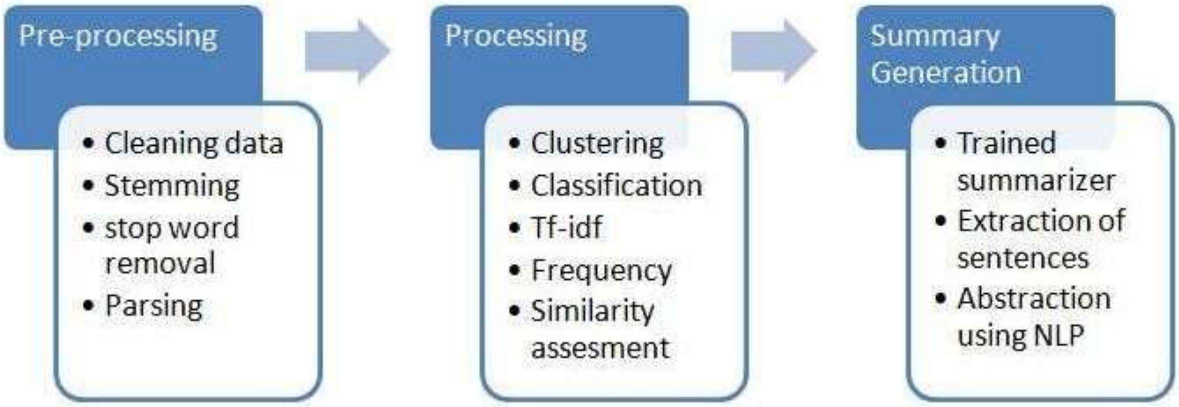
1) **Pre-processing,**

2) **Processing,**

3) **Summary Generation.**

Pre-processing

Pre-processing is defined here as cleaning the data. For cleaning unwanted characters, symbols, extra spaces, hyperlinks etc are removed. The stop words like ‘a’, ‘the’, ’and’. Stemming is done where the words are reduced to their word root for example ‘playing’ would be reduced to ‘play’. Moreover the parsing of the above data is done where the words are bifurcated into nouns, adjectives, verbs etc. The pre-processing is done as per the requirement using one or combination of the above defined techniques.



Processing

This is the next phase of document summarization. After the Text data is cleaned and pre-processed, the processing techniques are applied in which the tf-idf scores, frequency of words are calculated. The data is grouped on the basis of similarity, dissimilarity so that the summary generation can be made efficiently. For this different clustering techniques are used.

Summary Generation

After the data is processed, the summary is to be generated based on the requirement. Extraction of sentences is done from the processed data and the summary is processed. The words to added to sentences, reducing the sentences based on score etc is done in this step to produce summary. The representation of the summary can be in the form of words, sentences, paragraphs, graphs etc. for the summary to be generated different summarizers are used.

2.AUTOMATIC KEYWORD EXTRACTION: A REVIEW

On the premise of past work done towards automatic keyword extraction from the text for its summarization, extraction systems can be classified into four classes, namely, simple statistical approach, linguistics approach, machine learning approach, and hybrid approaches.

2.1 Simple Statistical Approach

These strategies are rough, simplistic and have a tendency to have no training sets. They concentrate on statistics got from non-linguistic features of the document, for example, the position of a word inside the document, the term frequency, and inverse document frequency. These insights are later used to build up a list of keywords. Cohen, utilized n-gram statistical data to discover the keyword inside the document automatically. Other techniques in- side this class incorporate word frequency, term frequency (TF) or term frequency inverse document frequency (TF-IDF) , word cooccurrences , and PAT-tree . The most essential of them is term frequency. In these strategies, the frequency of occurrence is the main criteria that choose whether a word is a keyword or not. It is extremely unrefined and tends to give very unseemly results. An improvement of this strategy is the TF-IDF, which also takes the frequency of occurrence of a word as the model to choose a keyword or not. Similarly, word co-occurrence methods manage statistical information about the number of times a word has happened and the number of times it has happened with another word. This statistical information is then used to compute support and confidence of the words. Apriori technique is then used to infer the keywords.

2.2 Linguistics Approach

This approach utilizes the linguistic features of the words for keyword detection and extraction in text documents. It incorporates the lexical analysis , syntactic analysis, discourse analysis , etc. The resources used for lexical analysis are an electronic dictionary, tree tagger, WordNet, ngrams, POS pattern, etc. Similarly, noun phrase (NP), chunks (Parsing) are used as resources for syntactic analysis.

2.3 Machine Learning Approach

Keyword extraction can also be seen as a learning problem. This approach requires manually annotated training data and training models. Hidden Markov model , support vector machine (SVM) , naive Bayes (NB) , bagging , etc. are commonly used training models in these approaches. In the second phase, the document whose keywords are to be extracted is given as inputs to the model, which then extracts the keywords that best fit the model’s training. One of the most famous algorithms in this approach is the keyword extraction algorithm (KEA) . In this approach, the article is first converted into a graph where each word is treated as a node, and whenever two words appear in the same sentence, the nodes are connected with an edge for each time they appear together. Then the number of edges connecting the vertices are converted into scores and are clustered accordingly. The cluster heads are treated as keywords. Bayesian algorithms use the Bayes classifier to classify the word into two categories: keyword or not a keyword depending on how it is trained. GenEx is another tool in this approach.

2.4 Hybrid Approach

These approaches combine the above two methods or use heuristics, such as position, length, layout feature of the words, HTML tags around the words, etc. These algorithms are designed to take the best features from above mentioned approaches. Based on the classification we bring a consolidated summary of previous studies on automatic keyword extraction . It discusses the approaches that are used for keyword extraction and various domains of dataset in which experiments are performed.

3.DOCUMENT SUMMARIZATION PROCESS: A REVIEW

Based on the literature, text summarization process can be characterized into five types, namely, based on the number of the document, based on summary usage, based on techniques, based on characteristics of summary as text and based on levels of linguistics process.

3.1 Single Document Text Summarization

In single document text summarization, it takes a single document as an input to perform summarization and produce a single output document . Thomas *et al*. designed a system for automatic keyword extraction for text summarization in single document e-Newspaper article. Marcu *et al.* developed a discourse-based summarizer that determines adequacy for summarizing texts for discourse- based methods in the domain of single news articles.

3.2 Multiple Document Text Summarization

In multiple documents text summarization, it takes numerous documents as an input to perform summarization and deliver a single output document . Mirroshandel *et al*. presents two different algorithms towards temporal relation based keyword extraction and text summarization in multi-document. The first algorithm was a weakly supervised machine learning approach for classification of temporal relations between events and the second algorithm was expectation maximization (EM) based unsupervised learning approach for temporal relation extraction. Min *et al*. used the information which is common to document sets belonging to a common category to improve the quality of automatically extracted content in multi-document summaries.

3.3 Query-based Text Summarization

In this summarization technique, a particular portion is utilized to extract the essential keyword from input document to make the summary of corresponding document. Fisher *et al*. developed a query- based summarization system that uses a log-linear model to classify each word in a sentence. It exploits the property of sentence ranking methods in which they consider neural query ranking and query-focused ranking. Dong *et al.* developed a query-based summarization that uses document ranking, time-sensitive queries and ranks recency sensitive queries as the features for text summarization.

3.4 Extractive Text Summarization

In this procedure, summarizer discovers more critical information (either words or sentences) from input document to make the summary of the corresponding document . In this process, it uses statistical and linguistic features of the sentences to decide the most relevant sentences in the given input document. Thomas *et al*. designed a model based extractive summarizer using machine learning and simple statistical method for keyword extraction from e-Newspaper article. Min *et al.* used freely available, open-source extractive summarization system, called SWING to summarize the text in multi-document. They used information which is common to document sets belonging to a common category as a feature and encapsulated the concept of category-specific importance (CSI). They showed that CSI is a valuable metric to aid sentence selection in extractive summarization tasks. Marcu *et al.* developed a discourse- based extractive summarizer that uses the rhetorical parsing algorithm to determine discourse structure of the text of given input, determine partial ordering on the elementary and parenthetical units of the text. Erkan *et al.* developed an extractive summarization environment. It consists of three steps: feature extractor, the feature vector, and reranker.

Features are Centroid, Position, Length Cutoff, SimWithFirst, LexPageRank, and QueryPhraseMatch. Alguliev *et al.* developed an unsupervised learning based extractive summarizer that optimizes three properties: relevance, redundancy, and length. It split documents into sentences and select salient sentences from the document. Aramaki *et al*. destined a supervised learning based extractive text summarizer that identifies the negative event and it also investigates what kind of information is helpful for negative event identification. An SVM classifier is used to distinguish negative events from other events.

3.5 Abstractive Text Summarization

In this procedure, a machine needs to comprehend the idea of all the input documents and then deliver summary with its particular sentences . It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. Brandow *et al.* developed an abstractive summarization system that analyses the statistical corpus and extracts the signature words from the corpus. Then it assigns the weight for all the signature words. Based on the extracted signature words, they assign the weight to the sentences and select few top weighted sentences as the summary. Daume *et al.* developed an abstractive summarization system that maps all the documents into database-like representation. Further, it classifies into four categories: a single person, single event, multiple event, and natural disaster. It generates a short headline using a set of predefined templates. It generates summaries by extracting sentences from the database.

3.6 Supervised Learning Based Text Summarization

This type of learning techniques used labeled dataset for training . Thomas et al. designed a system for automatic keyword extraction for text summarization using hidden Markov model. The learning process was supervised, it used human annotated keyword set to train the model. Mirroshandel *et al.* used a set of labeled dataset to train the system for the classification of temporal relations between events. Aramaki et al. destined a supervised learning based extractive text summarizer that identifies the negative event and also investigates what kind of information is helpful for negative event identification. An SVM classifier is used to distinguish negative events from other events.

3.7 Unsupervised Learning Based Text Summarization

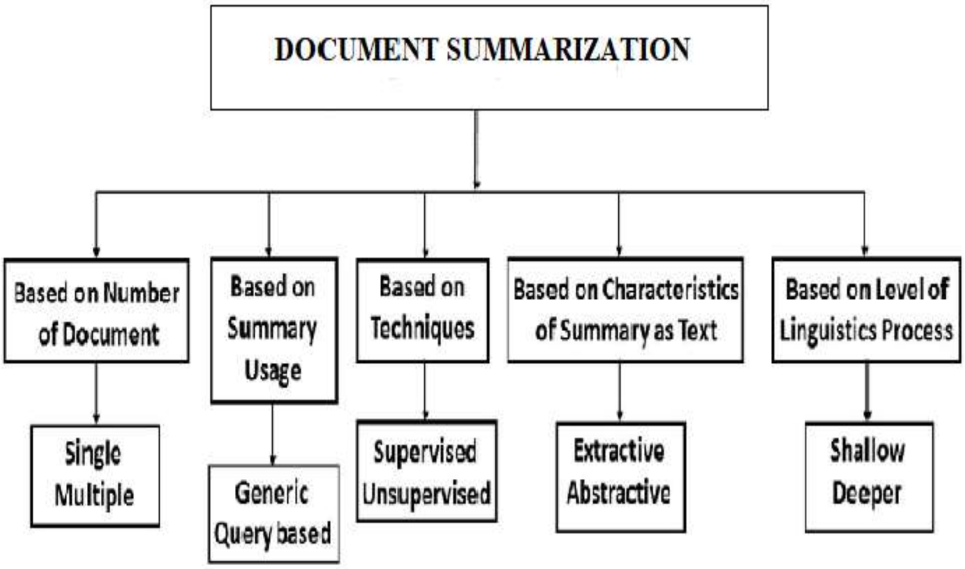


Figure 2: Characterization of the text summarization process

In this technique, there are no predefined guidelines available at the time of training . Mirroshandel *et al.* proposed a method for temporal relation extraction based on the Expectation-Maximization (EM) algorithm. Within EM, they used different techniques such as a greedy best-first search and integer linear programming for temporal inconsistency removal. The EM-based approach was a fully unsupervised temporal relation based extraction for text summarization. Alguliev *et al.* developed an unsupervised learning based extractive summarizer that optimizes three properties: relevance, redundancy, and length. It split documents into sentences select salient sentences from the document.

4.ISSUES AND CHALLENGES OCCURS IN TEXT SUMMARIZATION

In the area of text summarization, there are following research issues and challenges occurs during implementation.

Research Issues

- In the case of multi-document text summarization, several issues occurs frequently while evaluation of summary such as redundancy, temporal dimension, co-reference or sentence ordering, etc. which makes very difficult to achieve quality summary. Some other issues occurs such as grammaticality, cohesion, coherence which is harmful for summary.
- The quality of summaries are varying from system to system or person to person. Some person feels some set of sentences are important for summary, at the same time other person feel the other set of sentences are important for required summary.

Implementation Challenges

- To get the quality summary, quality keywords are required for text summarization.
- There is no standard to identify quality keywords within or multiple documents. The extracted keywords are varying for applying different approaches of keyword extraction.
- Multi-lingual text summarization is another challenging task.

5. CONCLUSION AND FUTURE ENHANCEMENTS

- This paper gives the brief information about document summarization.
- The summarization types, various stages in document summarization, different document summarization techniques are discussed in this paper.
- After the Literature review, papers which includes various techniques along with new and hybrid techniques of document summarization it can be seen that still there remain some drawbacks like , lower accuracy, no specific measure for how much the sentence is similar to other sentence due to which the computer generated summaries are not so efficient as the human generated summaries.

REVIEW TABLE

SL NO	TITLE	METHODLOGY	ALGORITHM	ACCURACY
1	A Neural Attention Model for Abstractive Sentence Summarization	local attention-based model	Beam Search	43.81%
2	Abstractive Text Summarization using Sequence-to-sequence RNN and Beyond	RNN	-	74.51%
3	Automatic Keyword Extraction for Text Summarization: survey	-	-	-
4	Summarization with Pointer-Generator Networks	Pointer generator network and CNN	-	-
5	Text Summarization Techniques: A Brief Survey	Frequency Driven Approaches	-	-
6	Ranking Sentences for Extractive Summarization with Reinforcement Learning	CNN/RNN	REINFORCE algorithm	37.98%
7	Neural Document Summarization by Jointly Learning to Score and Select Sentences	RNN	-	66.34%
8	Machine Learning Techniques for Document Summarization: A Survey	Talks about all the methods	-	-
9	Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization	R2N2/RNN	-	-
10	A Neural Attention Model for Sentence Summarization	local attention-based model	Beam Search	23.97%
11	A Review Paper on Text Summarization	Natural Language Processing	-	-
12	Automatic Text Summarization Using Natural Language Processing	Natural Language Processing	LESK Algorithm	-
13	A Novel Technique for Efficient Text Document Summarization as a Service	Natural Language Processing	LESK Algorithm	33%
14	Automatic Text Summarization and its methods : A Review	Term frequency based method/ Graph based method/ Time based method/Clustering based method	-	-
15	A Review on Automatic Text Summarization Approaches	Frequency Based Approach/Term Frequency–Inverse Document Frequency/Machine Learning Approach/Discourse Based Method	-	-

- Document summarization is very helpful for users to extract only needed information in stipulated time.
- In this area, considerable amount of work has been done in the recent past. Due to lack of information and standardization lot of research overlap is a common phenomenon.
- Since 2012, exhaustive review paper is not published on automatic keyword extraction and document summarization especially in Indian context.
- Therefore, we thought that, the survey paper covering recent work in keyword extraction and text summarization may ignite the research community for filling some important research gaps.
- This paper contains the literature review of recent work in text summarization from the point of views of automatic keyword extraction, text databases, summarization process, summarization methodologies and evaluation matrices.

In future, one can target following direction in the field of summarization

- Future work can be done in defining a similarity measure with a threshold value which can state how much a sentence is similar to other sentence instead of relative similarity measure, increasing the accuracy of sentence extraction, reducing the redundancy, effective clustering of documents.
- Text summarization in low resourced languages especially in Indian language context such as Telugu, Hindi, Kannada, Bengali,etc.
- This work can also be extended to multi-lingual text summarization.
- Multimedia summarization.

6. REFERENCES

[1] Alexander M., Rush Sumit Chopra, Jason Weston- A Neural Attention Model for Abstractive Sentence Summarization arXiv:1509.00685v2 [cs.CL] 3 Sep 2015

[2] Çaglar˘Gulçehre˙ Bing Xiang, Ramesh Nallapati, Bowen Zhou, Cicerodos Santos - Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, arXiv:1602.06023v5 [cs.CL] 26 Aug 2016

[3] Santosh Kumar Bharti, Korra Sathya Babu, Sanjay Kumar Jena - Automatic *Keyword* Extraction for Text Summarization: A Survey, National Institute of Technology, Rourkela, Odisha 769008 India e-mail@nitrkl.ac.in 08-February-2017

[4] Abigail See, Peter J. Liu, Christopher D. Manning - Get To The Point: Summarization with Pointer-Generator NetworksarXiv:1704.04368v2 [cs.CL] 25 Apr 2017

[5] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut - Text Summarization Techniques: A Brief Survey, arXiv:1707.02268v3 [cs.CL] 28 Jul 2017

[6] Shashi Narayan, Shay B. Cohen,Mirella Lapata- Ranking Sentences for Extractive Summarization with Reinforcement Learning arXiv:1802.08636v2 [cs.CL] 16 Apr 2018

[7] Qingyu Zhou[y], Nan Yang[z], Furu Wei[z], Shaohan Huang[z], Ming Zhou[z], Tiejun Zhao[y] [y]Harbin Institute of Technology, Neural Document Summarization by JointlyLearning to Score and Select SentencesarXiv:1807.02305v1 [cs.CL] 6 Jul 2018

[8] Feny Mehta - Machine Learning Techniques for Document Summarization: A Survey, 2016 IJEDR | Volume 4, Issue 2 | ISSN: 2321-9939

[9] Ziqiang Cao, Furu Wei,Li Dong,Sujian Li,Ming Zhou- Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence 2153

[10] Alexander M. Rush Sumit Chopra Jason Weston A Neural Attention Model for Sentence Summarization, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 379–389, Lisbon, Portugal, 17-21 September 2015. c 2015 Association for Computational Linguistics

[11] Deepali K. Gaikwad[1] and C. Namrata Mahender A Review Paper on Text Summarization, *International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016*

[12] Neelima Bhatia, Arunima jaiswal – Automatic text summarization and its methods- A review , 978-1-4673-8203/16/$31.00 IEEE 2016

[13] Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Sowmya Kamath S, A Novel Technique for Efficient Text Document Summarization as a Service 2013 Third International Conference on Advances in Computing and Communications, 978-0-7695-5033-6/13 $26.00 © 2013 IEEE, DOI 10.1109/ICACC.2013.17(13)

[14] Pratibha Devihosur1, Naseer R - Automatic Text Summarization Using Natural Language Processing, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 04 Issue: 08 | Aug -2017

[15] Yogan Jaya Kumar, Ong Sing Goh, Halizah Basiron, Ngo Hea Choon and Puspalata C Suppiah- A Review on Automatic Text Summarization Approaches, 2016 Yogan Jaya Kumar, Ong Sing Goh, Halizah Basiron, Ngo Hea Choon and Puspalata C Suppiah. This open access article is distributed under a Creative Commons Attribution (CC-BY) 3.0 license