Machine Learning Techniques for Document Summarization: A Survey

Feny Mehta
P.G Student of Computer Engineering
Marwadi Education Foundations Group of Institutions, Rajkot, India.

***Abstract* - Currently huge amount of data is available on the internet which is increasing exponentially day by day. It becomes time consuming and tedious job to search a specific topic from the heap of information available. Document summarization is the key solution to the above stated problem. It refers to reducing the size of the document still preserving the main information of it. Abstractive and Extractive are the two main automatic document summarization techniques. The aim of this paper is to present a survey on various extractive document summarization techniques.**

***Keywords* - Automatic Document summarization, hierarchical clustering, k-means clustering, sentence scoring, sentence extraction.**

I.INTRODUCTION

Due to the exponential growth of textual information available on the web, it becomes time consuming for the users to get desired information from the huge text corpus. If the users are provided with the summary of text available without losing its important information, it would be of a great help. This task can be performed with the help of document summarization Document summarization can be defined as representing a subset of data which contains the information of the entire set. Moreover automatic document summarization creates a summary or abstract of the entire document by selecting the most informative sentences from the document automatically through computer techniques or algorithms. Now Automatic Document summarization is a process that takes a source text and presents the most important content in a condensed form in a manner sensitive to the user or task needs.

In practice it is a tough task to generate an automatic summary because it involves deep natural language processing capacities and the system has to understand the point of a text. This requires semantic analysis and grouping of the content using word knowledge. Therefore, attempts at performing true abstraction have not been very successful so far. Now for summarizing the data, there is a need to group the similar information and then provide the summary of it. For example if numerous data is available for a single event occurring, the information common in each and important one can be extracted giving the summary and the main idea about the event, cutting down the extra information of it which is not required. There is no specific measure to determine how much the sentence is similar to other sentences in the document, hence only its relevant importance can be given.

There are various issues in summarizing the text content. It is a difficult task to cluster the documents as the number of documents is very large and the document vectors are high dimensional. The other issue in document summarization is the redundancy problem. Again the saliency of the document is a major issue for document summarization. The sentences serving same meaning but different words can have same weightage and can be included in the summary increasing redundancy. Sometimes words serving more importance rarely appear and they are not given weights which may lead to lower accuracy and irrelevance of summary to the original document. Hence it also gives rise to coverage problem for the information of the document. Although so much research has been done on automatic generation of summary, still the summaries generated by computers have not reached the satisfactory level as the summaries generated by humans. To overcome these issues many researches are done and still there is a tremendous need of rigorous research and development in the field of automatic document summarization.

This paper is structured as follows: Section 2 describes the types of document summarization, section 3 discusses about the basic steps for document summarization, section 4 gives the brief information about various techniques of document summarization, section 5 contains literature review of different research papers related to document summarization and at last section 6 holds conclusion and future scope in the field of automatic document summarization.

II.PHASES OF DOCUMENT SUMMARIZATION

There are basic three phases of document summarization as follows:

1) Pre-processing,
2) Processing,
3) Summary Generation.

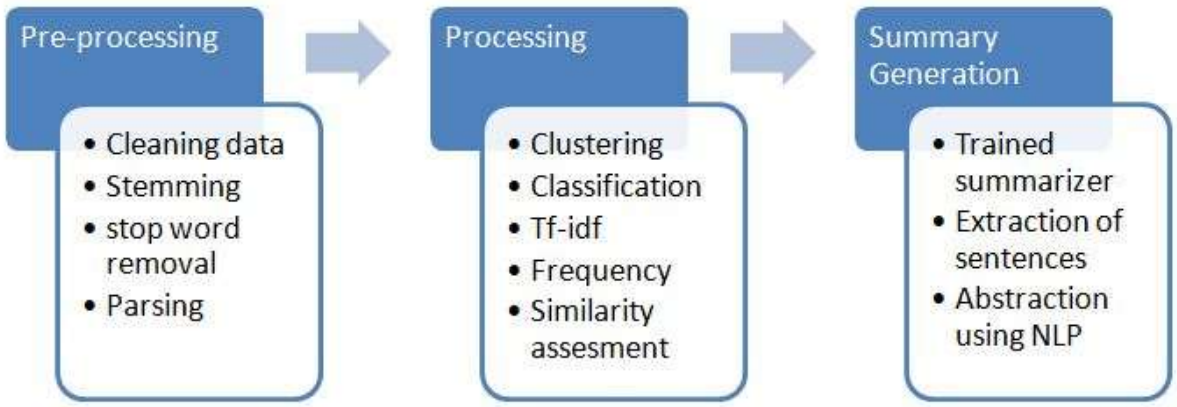

FIG 1: Phases of Document Summarization

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.

III. TYPES OF DOCUMENT SUMMARIZATION

Document summarization is mainly divided into following types:

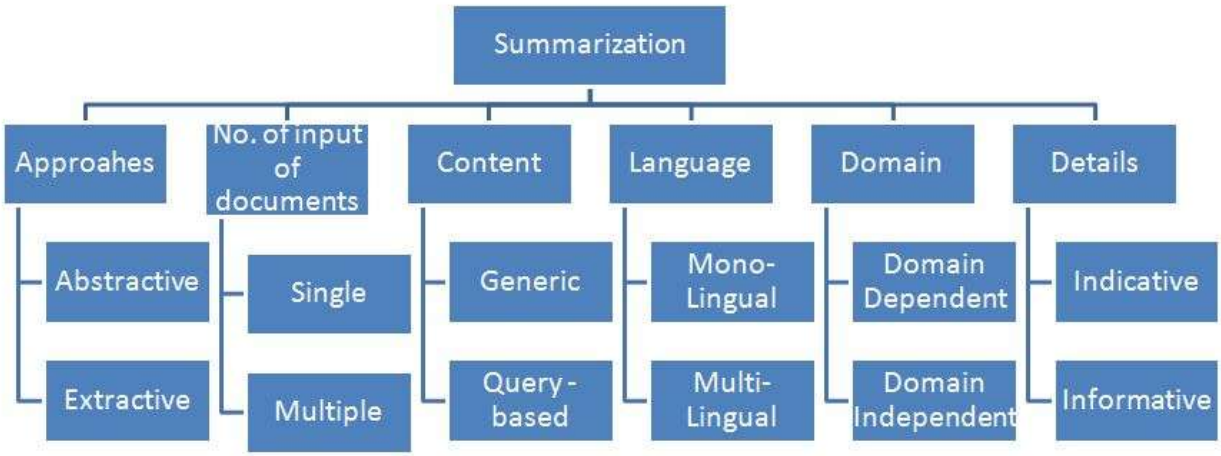

Fig 2: Types of Document summarization

Based on types of Approaches
There are two type of summarization based on their appraoches.
1) Abstractive method

Abstractive method builds an internal semantic representation and then uses Natural Language Processing (NLP) to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original.

2) Extractive method
In extractive summarization selection of a subset of existing words, phrases or sentences in the original text to form a summary is done.

Based on Number of Input Documents
1) Single document
The summary of a data from a single document is generated giving the brief idea about the document.
2) Multi document
The summary from two or more documents is given based on similarity or dissimilarity of the content in the documents.

Based on Content
1) Generic:
The generalised summary of the document is given based on the frequency and the importance of the sentence.
2) Query-based
The summary is generated based on the query from the user about a single topic, word etc.

Based on Language
1) Mono-lingual
The summary only for a single language can be generated and it may not give results for other languages.
2) Multi-lingual
The summarizer can produce summary for multiple languages .

Based on Domain
1) Domain Dependent
In this the summary to be produced is dependent on a specific domain and cannot be applied for different domains
2) Domain Independent
The summaries formed under this category does not depend on any specific domain and the result can be obtained for all the domains.

Based on Details
1) Indicative
In the indicative summaries the it mainly gives the main idea about the topic and does not give any other related information available in the document.
2) Informative
This type of summary gives the details related to the topic along with the important data hence giving more coverage.

IV.DOCUMENT SUMMARIZATION TECHNIQUES

Term frequency-Inverse document frequency
This technique is used for extraction of word to form the summary.
Tf-Idf of each document is generated using below formula:
$Tf\text{-}Idf_{t,d} = Tf_{t,d} * Idf_t$
Where Tf is term frequency of term t in document d and
$Idf_t = \log(N/df_t)$
Where N is the number of documents in corpus and
df is the document frequency of t.

1) Cluster based:
In this technique the sentences or words are clustered on the basis of similarity or dissimilarity and the data is extracted from each cluster to form the summary of that cluster. Different clustering techniques like k-means, hierarchical clustering etc are used

2) Graph theoretic :
In this technique the sentences or words are represented in the form of nodes of a unidirectional graph. This method gives the idea about the coverage about a specific sentence or word and also the most important once can be extracted on the basis of high cardinality.

4) Classification:
The data is classified into multiple classes as per the requirement. The data from the specific class can be extracted. Again this is a extraction technique for document summarization. Decision tree, rule based, support vector machine, k-nearest neighbour are various classification techniques used.

5) Neural network:
In neural networks, the information is processed in a similar way to te human brain. Initially it learns about all the features which must be present in the sentence, and then extracts them on the basis of it, removes dissimilar features and merges the similar ones.

6) Ontology based:
This is a abstraction based technique where the dictionary of words is used for the summarization process.

7) Fuzzy logic based:
In this summarization technique, instead of binary output the output is obtained in n-ary form.

8) Sematic based:
This summarization technique uses the semantic analysis abstraction method to generate summaries which are more accurate like the human generated summaries

V.LITRATURE REVIEW

Tanmay Basu et al. in [1] introduced a hybrid document clustering technique by combining a new hierarchical and the traditional k-means clustering techniques. A distance function is proposed to find the distance between the hierarchical clusters. Initially the algorithm constructs some clusters by the hierarchical clustering technique using the new distance function. Then k-means algorithm is performed by using the centroids of the hierarchical clusters to group the documents that are not included in the hierarchical clusters.

You Quyang et al. In [2] developed a novel sentence selection strategy that follows a progressive way to select the summary sentences. Methodology describes the subsuming relationship between two sentences. Then a progressive sentence selection strategy based on the subsuming relationship is applied, where word relations are identified, giving the coverage of it. The summary is made based on the scores of sentences. Redundancy is controlled by penalizing repetitive words.

Yogesh kumar Meena et al. in [3] proposed a feature priority based filtering method for summarization where sentences are filtered using tf-isf scores, named entities and proper nouns. After POS scores to the sentences, the sentences are arranged in the decreasing order of their scores; the first sentence is taken as it is for the summary at first position.

Sara Botelho Silveira et al. in [4] described an approach that uses lexical and semantic, both sentence reduction techniques. Summarization is done in two phases, first clustering by similarity and then clustering by keywords. The sentences are ordered on the basis of tf-idf scores. Sentence reduction is performed by removing specific sentential constructions conveying less relevant information to the summary. The three main algorithms proposed for it are main clause, blind removal, and best removal. The final step of this algorithm determines if the new reduced sentence can replace the former sentence based on the sentence score, finally providing improved summary.

Deepak P et al. in [5] introduced a method in which the interesting phrases are mined which are correlated to the query phrase but can be assumed independent to each other. The phrases are scored based on the conditional query. The outline of the disk-based and in-memory word-specific indexes is given which can be used by the algorithms to discover top-k phrases given the AND and OR queries.

Jialu Liu et al. in [6] gave a novel framework which extracts the quality phrases along with the segmenting the phrases from the text data. Initially the frequent phrases candidates are generated; the quality of it is estimated based on concordance and in-formativeness of it. Then the frequency of the phrases is rectified through phrasal segmentation; lastly the low rectified frequency phrases are filtered giving the quality phrases.

Mohamed Abdel Fattah in [7] proposes an approach that uses statistical tools to improve content selection in multi-document automatic text summarization. It uses the trainable summarizer, taking into account several features, which are then used in combination to construct text summarizer model. For final summary to be generated the hybrid model of maximum entropy, naive Bayesian classifier, support vector machine is proposed that ranks the sentences on the order of importance.

J. Yang in [8] proposed a design of a text summarizer consisting of three steps. First sentence segmentation is done based on cue phrases. After that the rhetorical relations are used and feature vectors are generated. The summarizer is trained to extract important segments using the three supervised machine learning algorithms individually, each generating a summary.

J. Mana-Lopez et al. in [9] presented the classification capacity of clustering techniques, along with the indicative extract about the contents of several sources by means of multi-document summarization techniques. Two kinds of summaries are provided, one covering the similarities of each cluster, second showing the particularities of each document with respect to the common topic in the cluster. The document multi-topic structure has been used in order to determine similarities and differences of topics in the cluster of documents.

Martha Mendoza et al. in [10] described a method for single document summarization. The sentences are extracted on the basis of its position in the document, relation of sentence with the title, sentence length, cohesion, coverage. Then the memetic algorithm is applied giving the optimised result using guided local search. The proposed algorithm is then compared with the other existing methods showing its outperformance.

Jingqiang Chen et al. in [11] designed a multi-document summarization system for scientific documents. Initially co-occurrence base of scientific terms is created. Now the document is taken, parsed to get noun phrases and bigrams and trigrams are created. They are clustered on the basis of frequent term set of common facts (i.e. co-occurring terms in citations). In order to summarize the data, scores are assigned based on count and weight of salient features. Selected sentences are reordered on the basis of publishing date.

Dragomir R. Radev et. Al in [12] presented a multi-document summarizer in which the relevant documents to the topic are selected and clustered on the basis cluster based relative utility, cross-sentence informational subsumption (CSIS) . The sentences are ranked on the basis of centroid value, positional value, and first sentence overlap summed up together. The summary is generated by extracting and re-ranking the sentences

Jian-Ping Mei et al. in [13] described a new multi-document summarization approach. In it initially the sentence similarity matrix is obtained computing similarity between every pair of sentences. After it the fuzzy clustering with weighted mediods is performed. An exemplar weight is calculated considering the closeness of the sentence and the prototype weight of the sentence in the same cluster. Then the position score of the sentence is calculated. Then to form a summary the scores of exemplar and position of each sentence are combined and ranked in the descending order.

Ramiz M. Aliguliyev in [14] described a document summarization process in which the sentences are clustered using the normalised Google distance (NGD) finding the dissimilarity between the sentences globally and locally. The estimation of the no of clusters is computed. A discrete differential evolution algorithm is used for the further clustering purpose. For sentence extraction the average weights of the sentences with respect to clusters is defined. Sentences are ranked in the reverse order of their scores and the top ranked sentences are selected for generating the summary.

J. Atkinson in [15] developed a multi-document summarization framework for web pages. After the data is pre-processed, the rhetorical role identification of sentences is done using the CRF classifier. After that the relevant sentences are extracted on the basis of occurrence of words in the document. The ranking is given to these extracted sentences and are grouped and rearranged on the basis of semantic similarity, producing the summary.

VI. CONCLUSION

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 of 15 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. 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.

VII. REFERENCES

[1] T. Basu and C. A. Murthy, "A similarity assessment technique for effective grouping of documents," Inf. Sci. (Ny)., vol. 311, pp. 149-162, 2015.
[2] Y. Ouyang, W. Li, R. Zhang, S. Li, and Q. Lu, "A progressive sentence selection strategy for document summarization," Inf. Process. Manag., vol. 49, no. 1, pp. 213-221, 2013.
[3] Y. K. Meena, D.Goplani, " Feature priority based sentence filtering method for extractive automatic text summarization," International Conference on Intelligent Computing, Communication and Convergence, Procedia Computer Science, Volume 48, pp. 728-734, 2015.
[4] Sara Botelho Silveria, Antonio Branco, "Sentence reduction algorithms to improve Multidocument summarization," 5th int. conf. on agents and artificial intelligence, pp. 261-276, 2014.
[5] Dey and D. Majumdar, "Fast Mining of Interesting Phrases from Subsets of Text Corpora," 17th int. conf. on Extending Database Technology, pp. 193-204, 2014.
[6] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han, "Mining Quality Phrases from Massive Text Corpora," Proc. 2015 ACM SIGMOD Int. Conf. Manag. Data, pp. 1729-1744, 2015.
[7] M. Fattah, "A hybrid machine learning model for multi-document summarization," Appl. Intell., vol. 40, no. 4, pp. 592-600, 2014.
[8] W. Chuang and J. Yang, "Extracting sentence segments for text summarization: a machine learning approach," Proc. 23rd Annu. Int. ACM , pp. 152-159, 2000.
[9] M. D. E. Buenaga, "Multidocument Summarization : An Added Value to Clustering in Interactive Retrieval," vol. 22, no. 2, pp. 215-241, 2004.
[10] M. Mendoza, S. Bonilla, C. Noguera, C. Cobos, and E. León, "Extractive single-document summarization based on genetic operators and guided local search," Expert Syst. Appl., vol. 41, no. 9, pp. 4158-4169, 2014.
[11] J. Chen and H. Zhuge, "Summarization of scientific documents by detecting common facts in citations," Futur. Gener. Comput. Syst., vol. 32, pp. 246-252, 2014
[12] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," Inf. Process. Manag., vol. 40, no. 6, pp. 919-938, 2004.
[13] J. P. Mei and L. Chen, "SumCR: A new subtopic-based extractive approach for text summarization," Knowl. Inf. Syst., vol. 31, pp. 527-545, 2012.

[14] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," Expert Syst. Appl., vol. 36, no. 4, pp. 7764-7772, 2009.
[15] J. Atkinson, R. Munoz, "Rhetorics-based multi-document summarization," Expert Systems with Application, vol. 40, pp. 4346-4352, 2013.
[16] H. Hashimi, A. Hafez, and H. Mathkour, "Selection criteria for text mining approaches," Comput. Human Behav., vol. 51, pp. 729-733, 2015.
[17] C. Aggarwal and C. Zhai, Mining text data, vol. 4, no. 2(63). 2012.
[18] M. W. Berry and M. Castellanos, "Survey of Text Mining : Clustering , Classification , and Retrieval , Second Edition," 2007.
[19] 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 Syst. Appl., vol. 40, no. 14, pp. 5755-5764, 2013.