Modern Multi-Document Text Summarization Techniques

Yash Asawa, Vignesh Balaji, Ishan Isaac Dey

Abstract: ***Text Summarization is the technique in which the source document is simplified, valuable information is distilled and an abridged version is produced. Over the last decade, the focus has shifted from single document to multi-document summarization and despite significant progress in the domain, challenges such as sentence ordering and fluency remain. In this paper, a thorough comparison of the several multi-document text summarization techniques such as Machine Learning based, Graph based, Game-Theory based and more has been presented. This paper in its entirety condenses and interprets the numerous approaches, merits and limitations of these techniques. The Benchmark datasets of this domain and their features have also been examined. This survey aims to distinguish the various summarization algorithms based on properties that prove to be valuable in the generation of highly consistent, rational, summaries with reduced redundancy and information richness. The conclusions presented by this paper can be utilized to identify the advantages of these papers which will help future researchers in their study of this domain and ensure the provision of important data for further analysis in a more systematic and comprehensive manner. With the aid of this paper, researchers can identify the areas that present some scope for improvement and thereafter come up with novel or possibly hybrid techniques in Multi-Document Summarization.***

Index Terms: Abstractive, Extractive, Multi-document summarization, Text Summarization

I. INTRODUCTION

For recovering data, people generally use the web, for example, Google, Bing, Yahoo etc. Since the amount of material on the web is evolving quickly, for clients it isn't simple to discover pertinent and fitting data according to the prerequisites. When a client transmits a query on an Internet search engine for information or data then the reaction in most of the occasions is a great many documents and the client needs to confront the repetitive assignment of finding the fitting data from this ocean of responses. This issue is known as "Data Overloading"[1].

The essential objective of various multi-document summarization techniques is to create summaries which provide extensive inclusion, less redundancy in the information and extensive consistency between sentences [2]. In other words, the important content is removed from each data source and at that point is re-structured to generate summaries for multiple documents.

Revised Manuscript Received on April 21, 2020.

*** Correspondence Author**

Saravanakumar Kandasamy, Vellore Institute of Technology, Vellore, Indiaksaravanakumar@vit.ac.in,

Yash Asawa,Student**,** Bachelor's degree, Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

Vignesh Balaji, Student**,** Bachelor's degree, Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

Ishan Dey, Student**,** Bachelor's degree, Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

A number of research studies have addressed multi-document text summarization over the last ten years but so far, only four survey papers [3][4][5][6] have been submitted on this overview.

Even though the papers have done a decent job in covering the approaches presented in our target domain, most of the novel techniques and models were presented after the survey papers were published.

Over the last few years, the summarization domain has developed beyond imagination because of all the new efficient models and approaches that have been published [6].

There are numerous notations for content summarization based on the number of input sources, how the summary is generated, the reason driving the generation of the summary, language in which the documents are presented to the system and their category. Some earlier works on this field have presented a Fuzzy System approach for the generation of textual summaries and they aimed at validating its performance by employing it for the assessment of text but it lagged behind because it couldn't match the results of a neural model completely driven by data.

Another paper presented a deep learning methodology for our problem. They proposed the use of vectors embedded to represent the hidden semantic relations between the words. The vectors were acquired from models that were already trained by the provision of extensive amounts of data. The requirement of an exhaustively inclusive corpus and extensive training posed to be an issue [8]. The usage of eight significant pre-established properties to calculate the scores for each sentence was proposed in another paper. Fuzzy Inference Systems were used for enhancing the value of a summary created by a general statistical approach. However there is no data and description available for the sets of rules that were used, except that they were manually generated [9].

In general, humans with the relevant expertise can use their intelligence and domain knowledge to model the systems. But, in a lot of cases, this can prove to be a very difficult task for humans and hence one of the early researches proposed an automatic modelling approach that utilizes data [10]. In a more recent study, a method which implemented a rule generation mechanism that incorporates expert knowledge was proposed, along with some properties gathered from data [11]. Another important issue in summarization of text is the optimization issue which led to a group of researchers proposing a nature inspired optimization approach which was a multi-criteria optimization model related to Artificial Bee Colony abbreviated as ABC [12].

An approach which condenses the redundant information from the summary using Differential Evolution that adapts based on the content was also presented as an optimization solution [13]. A novel work related to the graph domain proposed an unsupervised content co-ranking algorithm that infuses the word-sentence similarities with graph based ranking methods in such a manner that mutual influence would deliver the principal situation of sentences and words more precisely [14]. Different from previous research, another method used density peaks grouping summarization method that takes both redundancy and relevance removal by a single process [15]. A group of researchers recently used heuristics to optimize the classification precision by increasing the regions over which their cover was present [16]. Multi Document summarization across various languages has been just as interesting of a topic for researchers [17]. An incorrect word can cause several errors in the 86 translation, especially in its local context. Hence, the author proposed a framework to generate English-to-Chinese CLTS based on the translation metric of sentences [18].

A new idea led to development of Support Vector Regression abbreviated as SVR, which used the score from the quality of translation in the PageRank algorithm to identify the relevance between sentences for executing the generation of summaries across languages.[19]. Later, to improve the performance based on the usage of phrase-based translation framework, another researcher introduced phrase-based frameworks to parallely operate extraction, sentence scoring and compression of sentences [20].

In another paper an approach was proposed which uses both, the source data and the target language data to generate summaries across languages [21]. A researcher used data extraction and information mining methods to tackle the issues occurring during the summarization procedure [22]. And another method introduced graph-based approaches to surpass the challenges occurring during the summarization process [23]. In the graph-based document methods, the generated summary and its performance was not up to the mark. To overcome the limitation of these approaches, semantics-based summarization methods were used by a team of researchers to generate very precise summaries. Such strategies analysed the linguistic idea of the base document to produce semantically logical summaries [24]. To resolve the issue of generic summarization a centroid method that produces summaries based on inter-sentence and sentence-level features was used [25].

A method based on Hierarchical Dirichlet process was able to produce good results on a dataset that focussed on queries. HDP is a hybrid technique that obtains the relationship between sentences [26]. It was found that in the case of abstractive summarization, training from one end to the other by employing neural networks that depend on the encoder-decoder format has achieved significant success [27].

The first use of neural sequence-to-sequence method led to inaccurate text summarization. It used a corpus with CNN/Daily Mail data to train models in a supervised manner to generate summaries with multiple sentences by using individual documents [28]. More recently, it was proposed that summaries can be generated by getting rid of the rather unimportant words from the base document without substituting them which ensures the absence of any new words and guarantees the use of words that are present in the original document only. [29].

Recently, neural network-based techniques which are limited to extractive summarization are being used to address the shortcomings of having multiple documents as input and there are no obvious ways to control the redundancy in the summary [30].

Few researchers offered an explanation for the neural network techniques on summarization [31]. Another text summarization method claimed to generate an opinion-based summary review [32]. This technique is based on the public ontological knowledge base to condense opinions [33]. According to a research, a method was proposed, which produces a text summary by choosing and ranking [34]. Other studies also suggested a method to condense the user's analysis on Rice cooker [35].

A multi-text summarization method to summarize a user's analysis measured four main factors: ‘credibility’, ‘review time’, ‘review usefulness’ and ‘conflicting opinions [36]. A novel approach used a multi-objective optimization problem which was revealed to perform better than the single-objective case [37].

In the process of extractive summarization, the problem of generating the optimal summary can be structured as a combination optimization problem which is found to be NP-hard [38]. In fact, there was a method based on studies about the performance of the population-based stochastic exploration algorithms [39] according to which, the length restriction of the generated summary increases the performance time and reduces the proficiency of the algorithm [40]. The Multi Objective Artificial Bee Colony abbreviated as MOABC, has been proposed and obtained very good results [41].

The three important stages in sentiment analysis are: 1) sentence level, 2) document level and 3) Aspect level. A sentence or a document is classified as positive or negative or subjective or objective with regard to their polarity and subjectivity metrics [42]. A machine learning (ML) founded on the algorithm uses a list of better known machine learning approaches to order sentiment orientation [43]. The sentiment value of a sentence is calculated as the addition of the sentiment points of all the words in the sentence [44]. A text summarization technique was used to generate a brief opinion summary of reviews based on a deep learning approach and SentiWordNet based approach [45]. A sentiment summarization summarizes sentiments from a huge quantity of reviewers or multiple reviews [46].

The [47] problems that occurred when there is high topic multiplicity and redundancy in multi-document content summarization proposed a method that states that every summarization method should adhere to maximum coverage and relevancy and minimum redundancy [48]. In order to further enhance the performance of automatic text summarizer, it was [49] suggested using syntactic parsing by incorporating a multilayer approach that involves using dimensions such as degree and strength to compute the weightage for each node in a network of documents.

One research [50] claims that most summarisers that involved optimisation did not concentrate on increasing the fitness value and slowing down the convergence, which is why the recently released Shark Smell Optimisation (SSO) is being used in the proposed method as it does this

with the aid of gradient operation [51] and explains that stemming actually interferes with semantic analysis.
Abstractive MDS researches mostly include sentence fusion-based tactics [52] that involve grouping sentences and then choosing the most representative sentence from each group and then producing a new sentence for each bunch by extracting the common data of the cluster. Some information extraction-based methods which first remove important information units, such as sentence compression, phrases replacement and coreference resolution, are shown to produce more useful and concise summaries [54].
In recent years, [55] there have been a few summarization approaches trying to use approaches like shallow semantic parsing, such as semantic role classification and abstract meaning representation (AMR) to advance document summarization.

II. GENERIC DIAGRAM

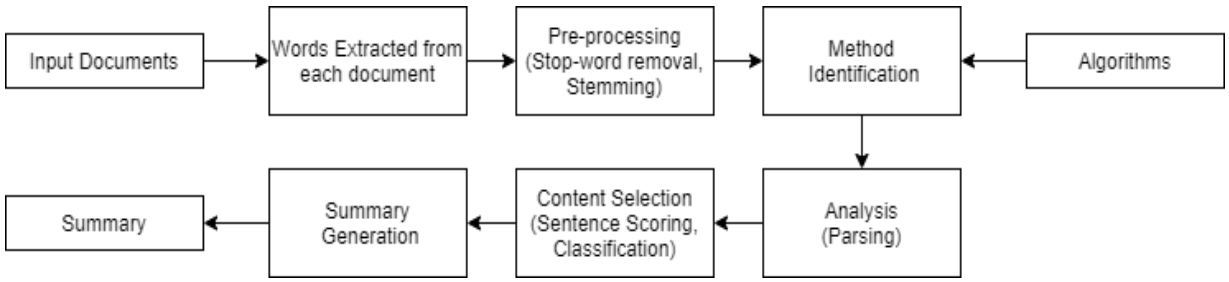

Figure 1: Diagram showing the working of Multi-Document Text Summarization

III. RELATED WORK

1. Abstractive Multi-Document Summarization

This approach involves the process of producing an eye-catching sentence or a short summary by constructing a few novel sentences that capture the noticeable notions of an article or a prose.

1.1. Neural Sentence fusion

.1.1. Introduction

The minds behind this approach [67] have proposed a method to improve the accuracy by using a set of related sentences as input for the encoder to check the risk of providing wrong details. They also developed new models for other NLP errands during their work for developing models to generate summaries. Utilizing transformers for the grouping task gave them somewhat better outcomes, however it is a rich model and it is computationally costly. Here, their primary concern is the MDS task, so they have not considered transformers for the grouping model to guarantee both time and proficiency.

1.1.2. Proposed Method

In order to address text clustering problems, the authors have used embedded words and have relied on a neural network with significant depth to ensure better rendition of text and therefore offer a clustering model for the sentences which is capable of getting trained in the absence of any supervision. They then use the Transformer model to achieve Neural sentence fusion. They have chosen the transformer model since it is an efficient and working process.

1.1.3. Evaluation of approach

Evaluation of the clustering behaviour is done based on accuracy (ACC) and the normal mutual information score (NMI). Normalization means computation of the Mutual Information (MI) metric to measure the outputs amid 0 - meaning no mutual information and 1 - meaning perfect correlation) between two clusters.

1.1.4. Advantages

1) Their model attains top clustering behaviour on ACC and NMI for the two datasets utilizing quick text word insertion.
2) Their model together improves the data coverage (BLEU, GMS) and complete abstraction (METEOR, Copy Rate, EACS) with a balanced compression ratio (CR).
3) Their system produced summary is very similar to man-made summaries in terms of Copy Rate and EACS metrics both.

1.1.5. Disadvantages

The key limitation of this method is that their neural framework for multi-document summarization only works for specified length which controls the shape of the summaries. In future, they will largely focus on contributing a superior neural architecture to encrypt a multi-document set. Also, they will try to propose a new sentence abstraction method (e.g., syntactic reorganization) using bi-directional beam search.

1.2. Semantic Link Network

1.2.1. Introduction

These authors [68] projected that ideas and occasions are essential for the semantic portrayal of records like news articles. Occasions show activities identified with ideas, ordinarily in type of "who did what to whom when and where", speaking to the general worries of peruses on the records like news. As a case of Semantically Linked Network, Semantically Linked System of ideas and occasions is reasonable for communicating with the semantics of archives which comprise news articles. Summarization from multiple documents can be actuated by changing the info records to a Semantically Linked Network of ideas and occasions, reducing the network to acquire a reduced and logically accurate network, and then modifying the network to generate an appropriate textual summary.

1.2.2. Proposed Method

The approach presents a clear Semantically Linked Network representation of data, which is feature rich and has multiple attributes that ensure better handling of documents for multiple purposes, mainly for generating summaries. The summary generated by the network must include the data of the actions and concepts which are of utmost importance, simultaneously continuing to remain consistent. They modelled the summary generation problem as a structure forecast problem that aims to achieve a trade-off between information selection, maintenance of consistency and generation of information which is accurate and complete.

1.2.3. Evaluation

The results are obtained with ROUGE Evaluation and Pyramid Methodology. To demonstrate the general performance of their system, they compare it with some extractive baselines and a
number of state-of-the-art abstractive standards.

1.2.3. Advantages

The merits of this approach are:
1) Core data of documents can be transformed into summaries more efficiently by condensing the network based on semantic features.
2) The events and ideas with significant co-relevance can be identified with ease and combined to merge similar semantic data to take care of redundancy, and collective related information from diverse sentences.
3) The network that is created based on semantic relations between the nodes can make the coherence and the information presented in the summary more intense.

1.2.4. Disadvantages

The main drawback of this method is that the presented approaches primarily depend on the use of statistics or syntax-based examination of the text instead of the semantic data of the documents.

2. Extractive Multi-Document Summarization

Extractive text summarization implicates the selection of expressions and sentences from the base text to make up the new summary. Procedures consist of ranking the relevance of phrases in order to select only those most significant to the implication of the source.

2.1. Artificial Bee Colony

2.1.1. Introduction

In this paper, the authors [69] have presented various structures to parallelize the MOABC calculation. For the parallelization, two irregular number generators and the various calendars given as per the OpenMP standard have been broken down and thought about. Then, an asynchronous parallel structure has been created based on the conduct of honey bees. The DUC datasets have been utilized in the tests, demonstrating the great aftereffects of their methodology from the two perspectives: computational execution and summary quality.

2.1.2. Proposed Method

None of the other papers have presented a similar approach and the examination of this approach has not been done as of yet. Their proposed approach was based on the creation of a structure based on the way honey bees go about their routine lives and then parallelizing that structure in an asynchronous manner.

2.1.3. Evaluation

The authors used ROUGE evaluation metrics and ROUGE-N, ROUGE-L scores for the evaluation of the proposed approach.

2.1.4. Advantages

The plus points of this approach are:
1) The improvement in the processing time allows much faster summary generation.
2) The proposed algorithm provides nearly 87 % efficiency and very high speeds.

2.1.5. Disadvantages

The limitation of this method is that this algorithm only increases the speedup time and does not provide any optimization solution for multi-document summarization.

2.2. Feature Based Summarization

2.2.1. Introduction

This approach [70] shows that pre-established features in the domain can be highly fruitful in the task of summary generation. In the proposed approach several pre-processing steps are imposed on raw texts, and a fruitful feature vector is produced, based on already established features in the domain.

2.2.2. Proposed Method

This proposed arrangement concentrated on extricating enlightening synopses from different reports utilizing normally utilized hand-made highlights from the writing. The main examination concentrated on the age of a feature vector. Also, a few blends of these highlights were analyzed and a shallow multiple layer perceptron and two distinctively demonstrated fuzzy deduction systems were utilized to extricate striking sentences from writings in the DUC dataset.

2.2.3. Advantages

1. The FS arrangements had the choice to effectively recognize summaries commendable and summaries contemptible sentences all the more appropriately.
2. The approach is suitable for handling a high-dimensional feature vector for the generation of a comprehensive set of rules.
3. The generated sets of rules go beyond the manually generated sets and ensure better coverage and a more generalized rendition.

2.2.4. Disadvantages

1. The proposed approach is based on a shallow neural network which can be outperformed by a deeper network as per studies, especially when the domain is highly complex and the feature space requires high-dimensional processing.
2.A very large number of rules need to be generated for giving out significantly accurate predictions and this makes the feature vector too complex for a system based on rules.
3. The usage of features created manually is restricted.

2.2.5. Evaluation

The authors evaluated the system for multiple purposes and in most cases, the acquired feature vector was successful in

generating a better, more accurate rendition and hence the systems' summarization capability achieves a highly sought benchmark and thus validates the proficiency of the approach.

2.3. Fuzzy Logic Based Summarization

2.3.1. Introduction

This approach [71] showed that there is a necessity for efficiently sieving and scavenging useful information from the Internet. Along with acquiring desired information there's a need for efficient content coverage with good information diversity. There is, furthermore, a significant scope for improvement in the performance of the present summarizers. The aim of the summarization method is to produce a precise and continuous synopsis of the provided text by ensuring the coverage of the most significant part of the contents and with minimum redundancy from various input sources.

2.3.2. Proposed Method

To produce the synopsis in the proposed method, the content is arranged as per highest to lowest value received from a fuzzy logic framework. Various sentences are selected depending on the acquired ratio of compression and likelihood metric with other sentences included in the summary. The similarity index is calculated and the content is arranged in the final summary in accordance with its positioning in the original document to retain cohesiveness. Cosine similarity metric is utilized to remove content with similar sentences from extracted important content to produce a final summary for multi-document summarization.

2.3.3. Advantages

1. The system is successful in tackling the main issue of redundant information in multi-document summarization to a great extent.
2. The proposed approach outperforms all the other existing systems.
3. None of the features that have been used in the proposed method depend on the language, making the approach very efficient.

2.3.4. Disadvantages

1. Multi -Document Summarization has relatively higher ambiguity than single document summarization which leads to inaccuracies in the results of the sentence scoring algorithm.
2. The sentence ordering issue in generating coherent summary.

2.3.5. Evaluation

All the outcomes of all features were compared with implied Multi-Document Summarization on the basis of Recall metric, Precision metric and F-measure values. We can observe that the performance of the summarization system is enhanced than all other approaches with respect to ROUGE-2 Recall metric. This implies that the proposed system has a high sentence coverage in the last summary. The proposed summary generation system has better ROUGE-4 scores based on all Recall, Precision and F-measure than other frameworks. The system is successful in tackling the main issue of redundant information in multi-document summarization to a great extent and outperforms all the other existent systems. Also, none of the novelties used in the proposed approach depend on the language, making it highly scalable.

3. Sentiment-Based Methods

Sentiment examination and summary generation for multiple documents uses natural language processing, textual analysis, statistics, machine learning systems and knowledge of linguistics to analyse, recognise and acquire information from documents.

3.1. Query-based and Opinion-oriented

3.1.1. Introduction

The team behind this research [72] have introduced a novel approach which generates opinion oriented summaries based on queries for multiple documents. The approach joins numerous sentiment lexicons to enhance the word inclusion limits of the person's vocabulary. A significant issue in case of a lexicon based methodology is the semantic distance between the previous polarity of a particular word displayed in the dictionary and the polarity of the word relevant to a particular reference. Besides, the kind of content can also influence the efficiency of the approach. Subsequently, to handle the previously mentioned difficulties, the approach incorporates numerous systems to change the orientation of the prior word while additionally thinking about the sort of content. The approach also uses the approach based on Semantic Sentiment to decide the estimation score of a word on the off chance that it is excluded from an sentiment lexicon.

3.1.2. Proposed Method

The proposed approach (QMOS) focuses on acquiring and reducing the sentences with their opinions that are relevant to the query given by the user. This issue is solved by using a combination of syntax related and semantic information in order to classify an increased number of sentences that are related to the query. A query involves minimal words which makes the recognition of the more salient sentences for generating answers to the user's query very difficult. Nevertheless, the presented approach makes use of a technique to expand the content word so that it can overcome that problem.

3.1.3. Advantages

1) Unlike other techniques, the proposed approach is capable of recognising the difference between the meaning of two sentences by employing a combination of semantic and syntactic data.
2) The proposed approach also takes polarity specific to the context and the types of content in sentiment analysis into consideration.

3.1.4. Disadvantages

The limitation of this approach is that The QMOS only considers opinion-oriented text data for multi-document summarization.

3.1.5. Evaluation

They used the ROUGE-N metric to assess the efficiency of the proposed approach.

3.2. Linguistic Knowledge Based and Sentiment Oriented

3.2.1. Introduction

The proposed approach implements the knowledge of sentiments in the analysis of the value of sentiment for each sentence so that it can be used as an attribute to group the sentences. An automatic system that can have the option to recognize emotional data, group client's judgements and generate a summary of the reviews is essential for the clients. No technique has previously employed a model to embed the words, information related to the sentiments, statistics and linguistic information for the generation of summary that is oriented towards sentiments which led the authors to the approach proposed in this paper.

3.2.2. Proposed Method

Their proposed method makes use of handlers for negations and but-clauses to check the orientation of the word prior sentiment. To achieve improved performance, they performed a thorough performance study using multiple methods for feature selection and classification to acquire features that held utmost importance and were successful in finding an efficient machine learning classifier, correspondingly. The method proposed in [73] is applied to three significantly different datasets validating its potential.

3.2.3. Advantages

1) The experimental outcomes prove that the use of sentiment classification methods that are based on SVM with the selection of features using the Information Gain technique surpass other approaches in terms of ARS values
2) The proposed approach goes one step ahead of the other approaches and is capable of using the CWE technique to classify words that are synonyms.

3.2.4. Disadvantages

The limitation of this method is that this approach does not take execution time into consideration. Therefore, a large text document could increase computation time significantly.

3.2.5. Evaluation

SVM turned out to be the best classification approach in terms of the ARS value. The authors also plan to evaluate the execution of the proposed approach on different datasets in the future.

3.3. Dual pattern-enhanced representations model

3.3.1. Introduction

The generation of summary which focuses on the user's query needs to strike the right balance between the importance of the query and the contextual topic so that the content requirements of the user are fulfilled in a way that the query is accurately solved and the necessary information also reaches the user while ensuring that the length of the generated snippet is within a certain limit. Since it tends to the issue regarding the overload of information corresponding to big data, automatic summary generation for multiple documents has an expanding incentive for different genuine applications. It abridges data from various documents that offer an unequivocal or verifiable fundamental theme. It assists clients rapidly getting the most applicable and significant data in huge ext data assortments. MDS can be sorted into conventional and query-based MDS. Generic MDS removes the significant substance from a document collection without utilizing any earlier information or extra data, while question centered MDS is intended to create a most common summary mirroring the dense data that is firmly identified with the hidden given query, which communicates the data need of the client.

3.3.2. Proposed Method

The authors came up with a Dual Pattern improved rendition model which focuses on the queries rather than focusing on the features like the graph-based approach. Their proposed approach employs a model that enhances the patterns and generates renditions that are semantically rich and fittingly selective for the data. They also incorporated a relevance model that identifies the relevance for the query with respect to the sentences based on the patterns.

With these renditions based on the patterns, their approach is successful in the amalgamation of multiple metrics for indication into a single integrated model for highly efficient multi-document summarization.

3.3.3. Advantages

1. The proposed approach achieves the right balance between the importance of the query and the contextual topic so that the content requirements of the user are fulfilled in a way that the query is accurately solved and the necessary information also reaches the user while ensuring that the length of the generated snippet is within a certain limit.
2.A novel weight assignment process is employed to levy a sort of fine to the generated patterns based on quality which are indifferent to the given query and to augment the representative power of the patterns that are relevant to the query.

3.3.4. Disadvantages

1. Does not completely solve the major problems in summary generation based on the query, is incapable of achieving complete coverage of the topics and reaching the desired level of query relevance.

3.3.5. Evaluation

The proposed approach was evaluated based on ROUGE-N metric and Paired t-test and the results show that the approach is successful in meeting the topic coverage requirements and the extraction of query relevant patterns.

4. MCRMR: Maximum coverage and relevancy with minimal redundancy

4.1. Shark Smell Optimization method

4.1.1. Introduction

The authors. [75] have introduced a novel idea for multi-document summarization viz. As it is becoming increasingly common to get access to large quantities of data on any topic that is of interest to you, it is also becoming more and more compulsory to have a summarization tool under your armour in order to understand and obtain the information that you require without getting overwhelmed by the immense display of all the data. There are many types of summarization. It is usually classified into abstractive and extractive summarization.

Extractive summarization effectively just extracts most of the most relevant content and produces the summary output. Abstractive summarization involves actually understanding the content and building new sentences to cover all the topics and then producing the summary output. This paper is proposing a method for multi-document extractive summarization.

4.1.2. Proposed Method

The suggested approach is called MCRMR which stands for Maximum Coverage and Relevancy and Minimal Redundancy. So basically, the summary output must have 3 features: maximum coverage, maximum relevancy and minimal redundancy. This is essentially an optimization problem. Before the word similarities are calculated, first, the multiple documents are pre-processed, which means they go through sentence separation and stemming (stemming may be skipped) and stop-word removal. After this, using word embeddings, which is a numeric value given to words, the similarity of the words is calculated using distance function. The calculation of word similarity is a linear combination of Words Mover's Distance and normalized Google distance. In this way, the redundant sentences are removed so that the least similar sentences are combined into a single document. These sentences are then scored on the ground of quantified content scoring text attributes, such as sentence position, sentence length, sentence similarity, sentence containing title words and numeric data, number of proper nouns, sentences with frequent words, and finally sentence significance. The optimal weight of each feature is generated using Shark Smell Optimization SSO.

4.1.3. Advantages

The advantage of this approach is This proposed solution is more beneficial and efficient because of the use of effective similarity function and the latest and correct metaheuristic approach. Judging the summary by coverage and non-redundancy features also improved the performance of the proposed method.

4.1.4. Disadvantages

The limitation of this approach is If the multiple documents are not that similar, then the coverage will be extremely high and the single document generation will be in fact merging all the documents together, which is just fruitless and a waste of computational power.

4.1.5. Evaluation

This method is then compared to six other similar approaches on the DUC-6 and DUC-7 datasets. The results are in the manner of precision, recall and F1 measures. It is shown that the suggested method has the best F1 measure or it produces the most similar summary to the reference summary out of all the approaches. This shows the results of the correctly chosen and updated similarity functions and metaheuristic approach. In the future, the authors plan to explore the neural network based similarity approach, without tampering with the performance of the current method.

5. Graph-Based Multi-Document Summarization

In this paper, the authors have [76] proposed a graph-based technique for sentence positioning that is seen as the most significant methodologies in this field. Nevertheless, the larger part of these methodologies depends on just one weighting plan and one positioning technique, which may cause a few impediments in their frameworks.

5.1 Introduction

Text summarization can be named generic summarization where frameworks utilize all the significant data of the info reports in the created summary or query centered summarization in which frameworks summarize only the data in the information archives which is identified with a specific client question. As the years progressed, numerous significant works with various methodologies were proposed to distinguish the significant sentences for extractive summarization, for example, regulated methodologies, Graph-based techniques for sentence ranking have the benefit of utilizing information drawn from the whole content in making ranking choices as opposed to relying just upon neighbourhood sentence data. Likewise, graph-based techniques are completely solo and rely just upon the content to be summarized without the requirement for any training information.

5.2. Proposed Method

Their work depends on two significant works on diagram biased strategies for content positioning; TextRank, LexRank. Graph based strategies for sentence scoring have demonstrated to be effective for both single-archive and multi-report outlines. Such methodologies don't include any complex phonetic handling of the content other than recognizing its sentences and words. They likewise have the upside of being completely solo and rely just upon the content to be outlined without the requirement for any preparation information.

5.3. Advantages

1.Created an enhanced weighting system by putting together various important measures that compute the likelihood between two records.
2.Designed a simple approach that pulls out the summaries through simple methods that do not need tough linguistic processing or named training data.

5.4. Disadvantages

The limitations of this method are:
1.Making the average ranking results using the harmonic mean obtained comparable metrics to the PageRank and did not produce the desired improvement.
2. Not enough participated weighting schemes and ranking methods taken into account for the combination process.

5.5. Evaluation

Utilizing the harmonic average in joining weighting designs to better the arithmetic mean and shows a nice improvement to the two grounds and many state of the art frameworks. They've produced an enhanced system by combining multiple valid values that compute the similarity between two contents and it is a simple approach that pulls away the summaries through simple ways that do not require tough linguistic processing or named training data. Not enough participating weighting schemes and ranking methods were taken into picture for the combination process and taking the average ranking scores utilizing the harmonic average obtained comparable results to the PageRank and did not give the desired improvement.

6. Cross Language Text Summarization

These researchers [77] presented a solution to the problem viz. The readers require access to an informative summary of the various sources in a language that they understand which covers all the important aspects and provides all the information to the consumer in the frame of a brief and to-the-point summary.

6.1 Introduction

Cross-Language Text Summarization (CLTS) is about studying a record in a language input to get its features and produce a small, informative and accurate summary of this content in a target language. The systems developed for CLTS can be distinguished, like the Text Summarization (TS) attribute, based on if they are extractive, compressive or abstractive. The extractive TS generates a summary by joining the most related sentences of the documents, the compressive TS produces a summary by removing irrelevant data of sentences and finally, the abstractive TS creates a summary with new content that is not really contained in the source records. Many of the state-of-the-art approaches for CLTS are of the extractive group. They differ on which way they compute sentence similarities and raise the risk that translation mistakes are introduced in the generated summary. Recent systems have used abstractive and compressive approaches to improve the informativeness and the grammatical quality of summaries. However, these approaches require certain resources for each external data or language and collection of numerous methods that stop the adaptability of these systems to produce summaries into other languages.

6.2. Proposed Method

The compressive TS approaches in monolingual study adapt sentence and multi-sentence compression systems for the French to English CLTS issue to simply keep the main data. An LSTM system is built to study a sentence and choose which words stay in the compression. They also use an ILP system to shorten similar sentences while studying both grammaticality and informativeness. Then they combine sentence and multi-sentence compressions to generate more useful cross-lingual summaries. They simplified their last MSC method to just aim on the cohesion of phrases and a list of keywords to show the way to the compressions with the core information of these clusters formed based on their similarity in the source and destination languages to get the gist of these documents. The extension of the other bunch of their approach relies on compression techniques of a single sentence by deletion of words. Still with the goal to produce more useful summaries, they extended their last SC approach by adding an attention method to compressing sentences which stand on its own during the grouping step required by the MSC phase.

6.3. Advantages

1. MSC method looks for the same information and produces a small summary with selected keywords that summarize the main information.
2. The MSC version enhances the informativeness of summaries by generating shorter reports with the main data, which enables the addition of other relevant information in the summaries.
3. MSC does not need a training corpus to generate compressions.

6.4. Disadvantages

1. The SC method compresses first sentences of news that normally describe the key idea of the news in a straight way. However, SC are applied here for all types of content, e.g. sentences with difficult syntactic structure or with many matters. The proposed approach generates poor results for these kinds of sentences.
2. Absence of an analysis of the impact of the semantic analysis in the generation of cross-lingual summaries.

6.5. Evaluation

The compressions made by MSC improved the informativeness of the SC version; however, SC and MSC summary has the combination of mistakes generated by MSC and SC, which lessened the grammatical quality of synopsis. The MSC system looks for the same data and produces a short compression with selected keywords that summarize the main data. It improved the detailness of summaries by generating shorter sentences with the key information, which enabled the sum of other meaningful information in the summaries and does not need a training corpus to generate compressions.

7. Game Theory-based Summarization

7.1. Introduction

In this approach, the authors have [78] proposed an approach to implement text summarization using tools that are derived from using game theory. Most of the existing approaches involved algorithms used to cluster sentences and then extract each cluster's most relevant sentence(s). Another very common approach was the graph-based approach which involved representing the condensed input as a graph consisting of nodes and their relationships, which somewhat improved semantic flow. However the generated summary in some cases was not acceptable enough to be a summary as they were not semantically coherent enough and it lacked speed and performance.

7.2. Proposed Method

Recent trends have introduced the concept of a more intelligent text summarization that can produce a more semantically coherent summary output by taking into account the linguistics of the text. This kind of text summarization also results in boosting the performance. Wikipedia, a well-known ontological knowledge base is used to represent sentences in a way that preserves semantics. The proposed method includes modelling the submodularity and then mapping it as a budgeted maximum coverage problem, in order to overcome current summarization issues.

7.3. Advantages

1. Ontological knowledge base helps to preserve the semantic and conceptual information of text.
2. Moreover, the adoption of submodularity property helps to reduce redundancy and produce a more concise summary.
3. The use of semantic associations among concepts help to reveal the shrouded relationships existing in the text.
4. Generating the summary according to the submodularity models help to create a more diversified yet semantically coherent output

7.4. Disadvantages

1. Submodularity modelling can base the summary on discriminative sentences, where it represents a certain group.
2. Classifying semantic relationships is quite complex as there is more than one way to describe each association. More such semantic relations need to be explored and included for summarization purposes.

7.5. Evaluation

In terms of testing the proposed method, the best results came from DUC02 and DUC04, when compared to other existing methods, especially with respect to R-2 recall results. And with respect to empirical results obtained from query-based summarization datasets, game theory-based summarization showed the most growth in improvement, when compared to the other approaches. The integration of the Wikipedia hierarchy of concepts into this summarization technique has helped to form the public ontological knowledge base and therefore bring up the performance level. On top of that, the submodularity modelling helped to improve uniqueness and help create a more diversified output. The use of these semantic associations among concepts help to reveal the shrouded relationships existing in the text, thus resulting in a more unique and semantically coherent summary. The idea of submodularity modelling can be taken even further and implemented to achieve comparative summarization The concept of comparative summarization is basically dividing the set of documents into groups and then selecting the most representative documents of each group where each document is maximally dissimilar to the other groups. This has a variety of applications such as comparing related topics, content sources or authors. Classifying semantic relationships is more difficult than it looks as there are multiple ways to describe each association. More such semantic relations need to be explored and included for summarization purposes.

IV. EVALUATION METHODS

The efficiency of the compared models was evaluated using (ROUGE-n) scores which is Recall-Oriented understudy for evaluation and is widely accepted as the official performance evaluation tool for text summary generation systems. It involves a group of metrics which compare the summary generated by the model against summary generated by human volunteers which can be considered to be the benchmark, thus these metrics eliminate the need for a manual comparison and provide an accurate idea of the efficiency of the model. We have compared the efficiency of the various models considered for the survey by using their ROUGE-n scores upon being subjected to the same datasets.

V. DATASETS

The survey mainly uses the DUC 2001-2007 datasets. Out of these, the DUC 2002 dataset comprises various print media documents on specific subjects in English. The entire dataset has documents corresponding to 59 such subjects and there are 567 documents in total. Each topic included in the dataset has two sentence extracts provided as well. These extracts were chosen by human volunteers which adds to the validity of the dataset for evaluation purposes.

The DUC 2004 dataset was also used which includes a lot of articles that cover multiple subjects. These articles have been grouped based on their subjects so they can be used for the generation of summaries by the proposed systems.

The DUC 2006 was also used by some of the approaches which includes fifty sets of documents with 25 documents in each set. These document sets are accompanied by four gold standard summaries corresponding to each set where 250 words is considered to be the appropriate size for the target summary.

Another dataset that was used in a few of the compared papers is the Multiling Pilot 2011, which includes data for multiple languages, wherein each language has been represented using ten topics and ten texts have been included for each topic.

VI. COMPARISON OF VARIOUS METHODS

A. Neural Sentence Fusion Vs Semantic Link Network

In realizing the neural sentence fusion approach the multi-document set acquired from the user per-say is put through a sentence clustering system where the set is broken down into multiple clusters which are individually passed on to a deep neural network and the combined to carry out abstractive sentence selection and generate the output summary thereafter.

In order to address text clustering problems, the authors used word embedding and neural network designs with significant depth for improved illustration of the data and hence offered a model capable of performing unsupervised clustering of the sentences. They then used the Transformer model to achieve Neural sentence fusion. They chose the Transformer model because of its efficiency and working process.

Contrary to this [68] presented an approach based on the use of a clear Semantically Linked Network to represent the document, which makes it highly efficient in handling the documents for various uses like document summarization. The summary generated by the Semantically Linked Network inculcated the most significant actions and concepts data, and remained semantically cogent .

They modelled the summarization of the SLN as a structure forecast problem that traded off among picking salient information, maintaining coherence, and turning over correct and complete information. The coded documents undergo concept and event extraction before the identification of relations within the document and the Framenet Corpus is used for the construction of the Semantically Linked Network.

The Semantically Linked Network thus formed after combining the extracted data is summarized inculcating the coherence constraints and this is used to produce the summarized version of the input data as required by the end-user.

B. Feature Based Vs Query Based

The query oriented method is based on using a mixture of information that describes the semantics and the syntax of the data to classify more query related sentences.

A query involves very few words. So, recognizing important sentences to answer a user's query using this little data can be considered as the chief problem.

Nevertheless, [72] employs a content word expansion technique to get rid of this problem. The primary focus in this approach is the query itself and it implements a graph-based ranking model for the summarization wherein a Statistical Similarity Calculation Approach and a Combination Model work in tandem to generate the summary.

On the other hand the feature based approach presented in [70] uses primitive hand-crafted features and generates summaries with the appropriate information efficiently covered from all the documents presented by the user. Firstly, a feature vector is generated. A variety of features are used for this purpose which are based on factors considered essential in the processing of natural languages and for further processing to produce accurate summaries.

In the latter phase, these features are combined in multiple ways before being used as the different layers of a multi-layer perceptron with insubstantial depth and two fuzzy inference systems modelled appropriately in accordance with requirement of the system to make it capable of acquiring the more important sentences from the available data in the DUC Dataset.

C. Machine Learning Approach Vs Game Theory Approach

The machine learning approach makes use of handlers for negations and but-clauses to check the orientation of the word prior sentiment.

To achieve improved performance, they performed a thorough performance study using multiple methods for feature selection and classification to acquire features that held utmost importance and were successful in finding an efficient machine learning classifier, correspondingly. The method proposed in [73] is applied to three significantly different datasets validating its potential.

The review text is pre-processed and feature extraction takes place wherein multiple features are collected for Feature Selection and for creating a Document vector as per the knowledge of the sentiments, the embedded words and Statistics and Linguistics. The Opinion Words and WordNet are also inculcated into this stage to ensure quality. The vectors are then passed onto a classification model for summary generation.

Contrary to this [78] presents an approach based on Game Theory where the sentences from the corresponding document are represented through semantic concepts of Wikipedia, which proves to be an extensive database with ontological information. An adaptable framework is proposed based primarily on game theory which makes the most of the submodularity that exists between the sentences of the best documents, to take care of the different issues posed by summarization and generates summaries. Multiple documents are concatenated and sentences from them are mapped in accordance with the Wikipedia Knowledge Base. These mapped sentences then undergo Submodularity Modelling and the sentences are scored. Post this, Sentence Selection takes place which leads to Summary Generation.

D. Graph Based Vs Dual Pattern Enhanced Representation

[76] presented a graph-based approach and their work depends on the most significant works on diagram-based strategies for sentence positioning; TextRank and LexRank. Graph based strategies for sentence ranking have demonstrated to be effective for both single-archive and multi-report outlines. Such methodologies don't include any complex phonetic handling of the content other than recognizing its sentences and words. They likewise have the upside of being completely solo and rely just upon the content to be outlined without the requirement for any preparation information.

The document cluster is pre-processed, sentence similarity is quantitatively acquired and the sentences are ranked before being selected to extract the summary. Meanwhile in 2019 some researchers [74] came up with a Dual Pattern improved rendition model which focuses on the queries rather than focusing on the features like the graph-based approach. Their proposed approach employs a model that enhances the patterns and generates renditions that are semantically rich and fittingly selective for the data. They also incorporated a relevance model that identifies the relevance for the query with respect to the sentences based on the patterns.

With these renditions based on the patterns, their approach is successful in the amalgamation of multiple metrics for indication into a single integrated model for highly efficient multi-document summarization.

The tables 1,2,3 and 4 present a more feature specific comparison between the various algorithms.

FEATURE	NSF	SLN
Similarity to Human references	**High (in terms of Copy Rate and EAC values)**	**Low**
Redundancy	**High**	**Low**
Paraphrasing	**No notion of paraphrasing**	**Modifies words based on statistics and syntax**

Table 1 - Neuralsentence Fusion Vs Semantic Link Network

FEATURE	FB	QB
Coverage	High	Low
Complexity	High	Low
Redundancy	Low	High (does not address redundancy issues)

Table 2 - Feature Based Vs Query Based

FEATURE	MLA	GTA
Performance Issues	Low (Better performance)	High (Not enough semantic relations)
Execution Time	High (As input becomes larger, computational time increases significantly)	Low
Redundancy	Low	High

Table 3 - Machine Learning Approach Vs Game Theory Approach

FEATURE	GB	DPER
Performance	High	Low
Coverage	High	Low
Redundancy	Low	High

Table 4 - Graph Based Vs Dual Pattern Enhanced Representation

Base Papers	Method	Dataset Used	Rouge 1	Rouge 2	Rouge SU4	Rouge L

Neural Sentence fusion	**Abstractive**	**DUC2004**	**41.92**	**12.22**	**15.59**	**-**
Fuzzy Logic Based Summarization	**Extractive**	**DUC2004**	**-**	**0.099**	**-**	**0.038**
Game theory-based	**Extractive/Abstractive**	**DUC2004**	**-**	**0.1052**	**0.1052**	**-**
MCRMR: Shark Smell Optimization	**Extractive/Abstractive**	**DUC2004**	**0.410**	**0.136 (DUC2004**	**-**	**-**
Graph-Based Summarization	**Extractive/Abstractive**	**DUC2004**	**0.393**	**0.09983**	**-**	**-**

Table 5 - Rouge Metric Comparison on DUC2004 Dataset

Base Papers	Method	Dataset Used	Rouge 1	Rouge 2	Rouge SU4	Rouge L
QMOS	**Extractive/Abstractive**	**DUC2006**	**0.4079**	**0.0824**	**-**	**-**
Semantic Link Network	**Abstractive**	**DUC2006**	**0.39017**	**0.11033**	**0.14844**	**-**
Dual pattern-enhanced representations model	**Extractive/Abstractive**	**DUC2006**	**0.40551**	**0.09228**	**0.14966**	**-**

Table 6 - Rouge Metric Comparison on DUC2006 Dataset

Table 5 has presented a comparison of the ROUGE metrics on the DUC2004 dataset for five systems and Table 6 has depicted a comparison of the ROUGE values on the DUC2006 dataset for three frameworks.

Base Papers	Method	Dataset Used	Rouge 1	Rouge 2	Rouge SU4	Rouge L

Parallelizing MOABC	**Extractive**	-	-	**0.389**	-	**0.581**
SOSML	**Extractive/Abstractive**	-	-	-	-	-
Feature Based Summarization	**Extractive**	**Cross validation of MLP**	**0.6629**	**0.5908**	-	**0.6674**
Cross Language Text Summarization	**Extractive/Compressive/ Abstractive**	**MultiLing Pilot 2011**	**0.4743**	**0.1639**	**0.1947**	-

Table 7 - Rouge Metric Comparison for performance across comparable custom Datasets

Summarization Method	Dataset Used	ARS Value
QMOS	**DUC2004**	**0.2452**
SOSML	**DUC2001_1**	**0.3859**

Table 8 - ARS value Comparison on DUC2004 and DUC 2001_1 Dataset

Table 7 presents the ROUGE scores across numerous datasets for four models and table 8 collates the ARS scores for QMOS and SOSML models.

VII. RESULTS SIGNIFICANCE

Just coming up with the proposed method and architecture for multi-document summarization is not enough. There are already numerous techniques of multi-document summarization being used in practice. Each paper must convey why their proposed method is significantly better (or worse) and rock solid evidence must be provided to justify. This is done by carrying out significance tests.

These tests allow the authors to compare their proposed method's performance with already existing approaches statistically, and allows them to correctly conclude that their method has truly performed better than the others, and it is not just an anomaly or coincidence. Most of the papers use t-test, as the sample size is usually small. If the p-values are less than 0.05 i.e. for 5% significance, then it can be said that the better performance metrics produced by the proposed approach are statistically significant, and have not occurred by chance.

Table 9 presents the results of various significance tests from the base papers for a comprehensive comparison.

Summarization Method	Significance Test	Method Compared and Value

QMOS	**P-values from Wilcoxon's signed rank test**	**Summarizers**	**0.012**
		CCNU	**0.004**
		PolyU	**0.002**
		NUS	**0.030**
SOSML	**P-values from Wilcoxon's signed rank test**	**OMSHR**	**0.012**
		LSVRS	**0.011**
		CHOS	**0.012**
		TSAD	**0.012**
Fuzzy Logic Based Summarization	**P-values from Paired t-test (95% Significance Level)**	**System 65**	**1.6e-2**
		System 104	**4e-3**
		System 19	**2e-3**
		System 44	**4e-3**
MCRMR_SSO	**P-values from Paired t-test**	**MCRMR_PSO**	**1.9e-3**
		OCDSum	**3.3e-3**
		MCMR_B&B	**4.4e-3**

		LexRank	**2.4e-5**
Dual pattern-enhanced representations model	**P-values from Paired t-test**	**NIST Baseline - Lead**	**4.20e-17**
		NIST Baseline - CLASSY04	**6.00e-05**

Table 9 - Comparison of the results for independent significance tests

VIII. CONCLUSIONS AND FUTURE DIRECTIONS

In this day and age, the monstrous amount of data that can be accessed due to the huge strides made in technology and especially the World Wide Web, or more commonly known as the Internet, is frankly quite intimidating. In light of this, there emerged a need for automatic text summarisers that could diminish the sheer volume of data without altering the overall meaning and including all the main topics, so that users could digest the information without feeling too overwhelmed. In this paper, we attempted to illuminate on the various state-of-the-art methods of automatic multi-document text summarization, as well as give insight to the workings going on in the background.

Neural Sentence Fusion was the first to examine changing neural frameworks for the task of sentence fusion. Their main model combines three vital methods namely diversity, importance and coverage under a chosen size limit. Query-based opinion-oriented method, associates multiple sentiment lexicons to magnify the sentiment dictionary coverage. It also advances word coverage limits and to define the sentiment value of a word if it is not composed in a sentiment grammar. Parallelizing Multi Objective Artificial Bee Colony (MOABC) algorithm addresses the issue of the execution time for the summarization of multiple documents. However, none of these approaches are capable of addressing redundancy issues that arise during summarization. The AMOABC algorithm provided is the first parallel growing method applied to the multi document text summarization issue. Machine learning-based Sentiment-Oriented Summarization of Multi-documents using Linguistic knowledge (SOSML) technique does not take execution time into consideration. Therefore, an attribute collection technique offers better results in terms of ARS metrics. Feature Based Summarization uses Fuzzy sets based on a feature graph obtained from an automatic feature selection process which should've been an efficient approach. However, this vector is largely complex for a rule biased system, since a huge quantity of fuzzy rules are named and then made to make sensible predictions and the usability of using hand-crafted attributes is rather insignificant. Fuzzy logic-based summarization, which explains that Cosine likeliness measure is highly efficient in tackling redundant information across multiple documents also seemed like a viable methodology. A language detection component can make it possible to upgrade the proposed system into a multi-lingual, multi-document TS system. This system is successful in tackling the main issue of redundant information in multi-document summarization to a great extent, outperforms all the other existent systems and all the attributes used in this method are language non-dependent making it highly scalable. Its limitation is that it fails at addressing the sentence ordering issue in generating a coherent summary. Cross Language Text Summarization (CLTS) method, which claims that the Tree biased SC methods can be used to shorten long and intense content in order to mitigate the poor accuracy of NN approaches for these genres of sentences guarantees that the study of content in the target language acts as a more vital character to produce worthy cross lingual summaries. Text Summarization, even today, doesn't have a perfectly optimum solution. This paper discusses the advantages and limitations of various existing methods in hopes that the future researchers can develop more efficient or hybrid approaches based on the results of the aforementioned methods.

For future direction, novel methods or a mixture of two or more algorithms can be designed using the aid of natural language processing and language methods, which can be used to produce better summaries for multi-documents.

ACKNOWLEDGMENTS

The paper was made possible because of inestimable inputs from everyone involved, directly or indirectly. We would first like to thank our guide, Associate Prof. Saravanakumar K who was highly instrumental in providing not only the required innovative base for the project but also crucial and constructive inputs that helped us reach the compiled paper. Our guide helped us improve our understanding in the domain of Natural Language Processing and we are very thankful for his support all throughout the project.

Finally, we would like to thank Vellore Institute of Technology, for providing us with flexible choices related to the selection, planning and execution of the project and for supporting our research and execution related to the project.

REFERENCES

1. Kan, M.-Y., Klavans, J.L.: Using librarian techniques in automatic text summarization for information retrieval. In: Proceedings of the 2nd ACMIIEEE-CS Joint Conference on Digital Libraries, pp. 36–45. ACM (2002)

2. Meena, Y.K., Jain, A., Gopalani, D.: Survey on graph and cluster-based approaches in multi-document text summarization. In: Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, pp. 1–5 (2014). doi:10.1109/ICRAIE.2014.6909126
3. Shah, C., & Jivani, A. (2016, August). Literature study on multi-document text summarization techniques. In *International Conference on Smart Trends for Information Technology and Computer Communications* (pp. 442-451). Springer, Singapore.
4. Kumar, Y. J., & Salim, N. (2012). Automatic multi document summarization approaches.
5. Sekine, S., &Nobata, C. (2003, May). A survey for multi-document summarization. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5* (pp. 65-72). Association for Computational Linguistics.
6. Meena, Y. K., Jain, A., &Gopalani, D. (2014, May). Survey on graph and cluster-based approaches in multi-document text summarization. In International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014) (pp. 1-5). IEEE.
7. F.B. Goularte, S.M. Nassar, R. Fileto, H. Saggion, A text summarization method based on fuzzy rules and applicable to automated assessment, Expert Syst. Appl. 115 (2019) 264–275, http://dx.doi.org/10.1016/j.eswa. 2018.07.047
8. H. Kobayashi, M. Noguchi, T. Yatsuka, Summarization based on embedding distributions, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1984–1989.
9. Suanmali, L., Salim, N., &Binwahlan, M. S. (2009). Fuzzy logic based method for improving text summarization. arXiv preprint arXiv:0906.4690.
10. L.-X. Wang, J. Mendel, Generating fuzzy rules by learning from examples, IEEE Trans. Syst. Man Cybern. 22 (6) (1992) 1414–1427
11. B. Mutlu, E.A. Sezer, M.A. Akcayol, End-to-end hierarchical fuzzy inference solution, in: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2018, pp. 1–9
12. Sanchez-Gomez, J. M., Vega-Rodríguez, M. A., & Pérez, C. J. (2017). Extractive multidocument text summarization using a multi-objective artificial bee colony optimization approach. Knowledge-Based Systems.
13. Alguliev, R. M., Aliguliyev, R. M., &Mehdiyev, C. A. (2011). Sentence selection for generic document summarization using an adaptive differential evolution algorithm. Swarm and Evolutionary Computation, 1(4), 213–222.
14. Fang, C., Mu, D., Deng, Z., & Wu, Z. (2017). Word-sentence co-ranking for automatic extractive text summarization. Expert Systems with Applications, 72, 189–195.
15. Chen, K. Y., Liu, S. H., Chen, B., & Wang, H. M. (2018). An information distillation framework for extractive summarization. IEEE/ACM Transactions on Audio Speech and Language Processing, 26(1), 161–170
16. Liu, C., Wang, W., Tu, G., Xiang, Y., Wang, S., & Lv, F. (2017). A new centroid-based classification model for text categorization. Knowledge-Based Systems, 136, 15– 26
17. A. Leuski, C.-Y. Lin, L. Zhou, U. Germann, F. J. Och, E. Hovy, Cross643 lingual C*ST*RD: English Access to Hindi Information 2 (2003) 245–269.
18. X. Wan, H. Li, J. Xiao, Cross-language document summarization based 651 on machine translation quality prediction, in: Proceedings of the 48th 652 Annual Meeting of the Association for Computational Linguistics, ACL 653 '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 654 2010, pp. 917–926.
19. F. Boudin, S. Huet, J. Torres-Moreno, A graph-based approach to cross656 language multi-document summarization, Polibits 43 (2011) 113–118
20. J.-g. Yao, X. Wan, J. Xiao, Phrase-based compressive cross-language sum595 marization, in: Proceedings of the 2015 Conference on Empirical Methods 596 in Natural Language Processing, Association for Computational Linguis597 tics, 2015, pp. 118–127
21. X. Wan, Using bilingual information for cross-language document summa658 rization, in: ACL, The Association for Computer Linguistics, 2011, pp. 659 1546–1555.
22. Wang, S.; Li, W.; Deng, H.: Document update summarization using incremental hierarchical clustering. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 279–288 (2010)
23. Wang, D.; Zhu, S.; Li, T.; Chi, Y.; Gong, Y.: Integrating document clustering and multi-document summarization. ACM Trans. Knowl. Discov. Data 5, 14:1–14:26 (2011)
24. Conroy, J.M.; Schlesinger, J.D.; Goldstein, J.; O'Leary, D.P.: Leftbrain/right-brain multi-document summarization. In: DUC 2004 Conference Proceedings (2004)
25. Radev, D.R.; Jing, H.; Sty´s, M.; Tam, D.: Centroid-based summarization of multiple documents. Inf. Process. Manag. 40(6), 919–938 (2004)
26. Yang, C.C.; Wang, F.L.: Hierarchical summarization of large documents. J. Am. Soc. Inf. Sci. Technol. 59(6), 887–902 (2008)
27. Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: Proceedings of the ICLR 2015.
28. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P., 2015. Teaching machines to read and comprehend. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. MIT Press, Cambridge, MA, USA, pp. 1693–1701.
29. Narayan, S., Cohen, S.B., Lapata, M., 2018. Ranking sentences for extractive summarization with reinforcement learning. In: Proceedings of the NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics.
30. Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan, K., Radev, D., 2017. Graph-based neural multi-document summarization. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Association for Computational Linguistics, Vancouver, Canada, pp. 452–462.
31. Zhang, J., Tan, J., & Wan, X. (2018). Towards a neural network approach to abstractive multi-document summarization. arXiv preprint arXiv:1804.09010.
32. Lloret, E., Boldrini, E., Vodolazova, T., Martínez-Barco, P., Muñoz, R., & Palomar, M. (2015). A novel concept-level approach for ultra-concise opinion summarization, Expert Systems with Applications, 42, 7148–7156.
33. Lu, Y., Duan, H., Wang, H., &Zhai, C. (2010). Exploiting structured ontology to organize scattered online opinions. Proceedings of the 23rd international conference on computational linguistics (pp. 734–742). Association for Computational Linguistics.
34. Nishikawa, H., Hasegawa, T., Matsuo, Y., &Kikui, G. (2010). Opinion summarization with integer linear programming formulation for sentence extraction and ordering. Proceedings of the 23rd international conference on computational linguistics: Posters (pp. 910–918). Association for Computational Linguistics.
35. Wang, D., Zhu, S., & Li, T. (2013). SumView: A Web-based engine for summarizing product reviews and customer opinions. Expert Systems with Applications, 40, 27–33.
36. Hu, Y.-H., Chen, Y.-L., & Chou, H.-L. (2017). Opinion mining from online hotel reviews–A text summarization approach. Information Processing & Management, 53, 436–449.
37. H.H. Saleh, N.J. Kadhim, B.A. Attea, A genetic based optimization model for extractive multi-document text summarization, Iraqi J. Sci. 56 (2) (2015) 1489–1498.
38. R.M. Alguliev, R.M. Aliguliyev, M.S. Hajirahimova, C.A. Mehdiyev, MCMR: Maximum coverage and minimum redundant text summarization model, Expert Syst. Appl. 38 (12) (2011) 14514–14522.
39. R.M. Alguliev, R.M. Aliguliyev, M.S. Hajirahimova, Gendocsum+MCLR: Generic document summarization based on maximum coverage and less redundancy, Expert Syst. Appl. 39 (16) (2012) 12460–12473.
40. R.M. Alguliev, R.M. Aliguliyev, C.A. Mehdiyev, Psum-sade: a modified pmedian problem and self-adaptive differential evolution algorithm for text summarization, Appl. Comput. Intell. Soft Comput. 2011 (351498) (2011) 1–13.
41. J.M. Sanchez-Gomez, M.A. Vega-Rodríguez, C.J. Pérez, Extractive multidocument text summarization using a multi-objective artificial bee colony optimization approach, Knowl.-Based Syst. 159 (2018) 1–8.
42. Rana, T. A., & Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: comparative analysis and survey. Artificial Intelligence Review, 46, 459-483.
43. Lee, S., &Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. Expert Systems with Applications, 41, 3041-3046.
44. N., & Chatterjee, N. (2016). Text Summarization Using Sentiment Analysis for DUC Data. In Information Technology (ICIT), 2016 International Conference on (pp. 229-234): IEEE.

45. Raut, V. B., &Londhe, D. (2014). Opinion mining and summarization of hotel reviews. In Computational Intelligence and Communication Networks (CICN), 2014 International Conference on (pp. 556-559): IEEE.
46. Lloret, E., Boldrini, E., Vodolazova, T., Martínez-Barco, P., Muñoz, R., & Palomar, M. (2015). A novel concept-level approach for ultra-concise opinion summarization. Expert Systems with Applications, 42, 7148-7156.
47. Oufaida, H. ,Nouali, O. , &Blache, P. (2014). Minimum redundancy and maximum relevance for single and multi-document arabic text summarization. Journal of King Saud University-Computer and Information Sciences, 26 (4), 450–461.
48. Kusner, M. , Sun, Y. , Kolkin, N. , & Weinberger, K. (2015). From word embeddings to document distances. In International conference on machine learning (pp. 957–966) .
49. Tohalino, J. V. ,&Amancio, D. R. (2018). Extractive multi-document summarization using multilayer networks. Physica A: Statistical Mechanics and its Applications, 503 , 526–539 .
50. Abedinia, O. ,Amjady, N. , &Ghasemi, A. (2016). A new metaheuristic algorithm based on shark smell optimization. Complexity, 21 (5), 97–116 .
51. Toman, M., Tesar, R., &Jezek, K. (2006). Influence of word normalization on text classification. Proceedings of InSciT, 4, 354–358.
52. S. Banerjee, P. Mitra, and K. Sugiyama, "Multi-document abstractive summarization using ilp based multi-sentence compression". In IJCAI 2015, 1208– 1214.
53. P. E. Genest, and G. Lapalme, "Framework for abstractive summarization using text-to-text generation", In Proceedings of the Workshop on Monolingual Text-To-Text Generation, 2011, 64-73.
54. T. Cohn and M. Lapata, "Sentence compression as tree transduction", Journal of Artificial Intelligence Research, 2009, 637-674.
55. L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U.Hermjakob, and N. Schneider, "Abstract meaning representation for sembanking", In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (pp. 178-186), 2013.
56. S. Yan, and X. Wan, "SRRank: leveraging semantic roles for extractive multi-document summarization", IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2014, 22(12), 2048-2058.
57. J. Kupiec, J. Pedersen, and F. Chen, ``A trainable document summarizer," in Proc. 18th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 1995, pp. 6873.
58. J. Cheng and M. Lapata, ``Neural summarization by extracting sentences and words," 2016, arXiv:1603.07252. [Online]. Available: https://arxiv.org/abs/1603.07252
59. J. L. Neto, A. D. Santos, C. A. Kaestner, and A. A. Freitas, ``Document clustering and text summarization," in Proc. 4th Int. Conf. Practical Appl. Knowl. Discovery Data Mining, 2000, pp. 41-55.
60. R. Mihalcea and P. Tarau, ``Textrank: Bringing order into text," in Proc. Conf. Empirical Methods Natural Lang. Process., 2004, pp. 404-411
61. S. Brin and L. Page, ``The anatomy of a large-scale hypertextual Web search engine," Comput. Netw. ISDN Syst., vol. 30, nos. 1-7, pp. 107-117, Apr. 1998. doi: 10.1016/S0169-7552(98)00110-X.
62. J.M. Sanchez-Gomez, M.A. Vega-Rodríguez, C.J. Pérez, Extractive multidocument text summarization using a multi-objective artificial
63. bee colony optimization approach, Knowl.-Based Syst. 159 (2018) 1–8, http://dx.doi.org/10.1016/j.knosys.2017.11.029.
64. Y. Guangbing, W. Dunwei, Kinshuk, C. Nian-Shing, S. Erkki, A novel contextual topic model for multi-document summarization, Expert Syst. Appl. 42 (3) (2015) 1340–1352, http://dx.doi.org/10.1016/j.eswa.2014.09.015.
65. Y. Gao, Y. Xu, Y. Li, Pattern-based topics for document modelling in information filtering, IEEE Trans. Knowl. Data Eng. 27 (6) (2015) 1629–1642, http: //dx.doi.org/10.1109/TKDE.2014.2384497
66. E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, Multi-document summarization exploiting frequent itemsets, in: Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC '12, ACM, New York, NY, USA, 2012, pp. 782–786, http://dx.doi.org/10.1145/2245276.2245427.
67. J.P. Qiang, P. Chen, W. Ding, F. Xie, X. Wu, Multi-document summarization using closed patterns, Knowl.-Based Syst. 99 (2016) 28–38, http://dx.doi.org/ 10.1016/j.knosys.2016.01.030.
68. Fuad, T. A., Nayeem, M. T., Mahmud, A., &Chali, Y. (2019). Neural sentence fusion for diversity driven abstractive multi-document summarization. Computer Speech & Language, 58, 216-230.
69. Li, W., &Zhuge, H. (2019). Abstractive Multi-Document Summarization based on Semantic Link Network. IEEE Transactions on Knowledge and Data Engineering.
70. Sanchez-Gomez, J. M., Vega-Rodríguez, M. A., & Pérez, C. J. (2019). Parallelizing a multi-objective optimization approach for extractive multi-document text summarization. Journal of Parallel and Distributed Computing, 134, 166-179.
71. Mutlu, B., Sezer, E. A., &Akcayol, M. A. (2019). Multi-document extractive text summarization: A comparative assessment on features. Knowledge-Based Systems, 183, 104848.
72. Patel, D. B., Shah, S., &Chhinkaniwala, H. R. (2019). Fuzzy logic based multi Document Summarization with improved sentence scoring and redundancy removal technique. Expert Systems with Applications.
73. Abdi, A., Shamsuddin, S. M., &Aliguliyev, R. M. (2018). QMOS: Query-based multi-documents opinion-oriented summarization. Information Processing & Management, 54(2), 318-338.
74. Abdi, A., Shamsuddin, S. M., Hasan, S., &Piran, J. (2018). Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment. Expert Systems with Applications, 109, 66-85.
75. Wu, Y., Li, Y., & Xu, Y. (2019). Dual pattern-enhanced representations model for query-focused multi-document summarization. Knowledge-Based Systems, 163, 736-748.
76. Verma, P., & Om, H. (2019). MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. Expert Systems with Applications, 120, 43-56.
77. Alzuhair, A., & Al-Dhelaan, M. (2019). An Approach for Combining Multiple Weighting Schemes and Ranking Methods in Graph-Based Multi-Document Summarization. IEEE Access, 7, 120375-120386.
78. Pontes, E. L., Huet, S., Torres-Moreno, J. M., &Linhares, A. C. (2019). Compressive approaches for cross language multi-document summarization. Data & Knowledge Engineering, 101763.
79. Ahmad, A., & Ahmad, T. (2019). A Game Theory Approach for Multi-document Summarization. Arabian Journal for Science and Engineering, 44(4), 3655-3667.

AUTHORS PROFILE

Yash Asawa is currently pursuing his Bachelor's degree in Computer Science and Engineering from Vellore Institute of Technology, Vellore, India. His research interests include: Machine Learning, Artificial Intelligence, Natural Language Processing and Automated Systems.
Email – yash17bce2296@gmail.com
Address – Q-1208, Mens Hostel, Vit University, Vellore, India - 632014

Vignesh Balaji is currently pursuing his Bachelor's degree in Computer Science and Engineering from Vellore Institute of Technology, Vellore, India. His research interests include: Data Visualization, Artificial Intelligence and Natural Language Processing.
Email – ubvignesh04@gmail.com
Address – Mens Hostel, Vit University, Vellore, India – 632014

Ishan Dey is currently pursuing his Bachelor's degree in Computer Science and Engineering from Vellore Institute of Technology, Vellore, India. His research interests include: Machine Learning, Artificial Intelligence, Natural Language Processing and Full Stack Development.
Email – ishanisaac.dey2017@vitstudent.ac.in
Address – Mens Hostel, Vit University, Vellore, India – 632014
Saravanakumar Kandasamy, Vellore Institute of Technology, Vellore, India**ksaravanakumar@vit.ac.in,**