Volume 64, Issue 1, 2020

Journal of Scientific Research

Institute of Science,

Banaras Hindu University, Varanasi, India.

A Review on Text Summarization Techniques

Pradeepika Verma*[1] and Anshul Verma[2]

*[1]Dept. of CSE, Indian Institute of Technology (ISM) Dhanbad. pradeepikav.verma093@gmail.com

[2]Dept. of CS, Banaras Hindu University, Varanasi. anshulverma87@gmail.com

***Abstract*—In recent years, an enormous amount of text data from diversified sources has been emerged day-by-day. This huge amount of data carries essential information and knowledge that needs to be effectively summarized to be useful. Hence, the main contribution of this paper is twofold. We first introduce some concepts related to extractive text summarization and then provide a systematic analysis of various text summarization techniques. In particular, some challenges in extractive summarization of single as well as multiple documents are introduced. The problems focus on the textual assessment and similarity measurement between the text documents are addressed. The challenges discussed are generic and applicable to every possible scenario in text summarization. Then, existing state-of-the-art of extractive summarization techniques are discussed that focus on the identified challenges.**

***Index Terms*—Summarization, Graph based summarization, Meta-heuristic based summarization, Maximal marginal relevance based summarization, Evaluation.**

I. Introduction

Text summarization is an strenuous problem of Natural Language Processing (NLP) due to difficulty in interpreting every point of the text in a document. This requires a precise analysis of the text in various steps such as semantic analysis, lexical relations, named entity recognition, etc., which can be accomplished with a great deal of word knowledge only. Since it is hard to obtain the word knowledge in various aspects such as meaning of a word with respect to other content, related words, inferential interpretation, sentence generation, etc., generating abstracts as summaries have become complex. This type of summarization is classified as abstractive summarization in NLP. However, an approximation, which is classified as extractive summarization, is more flexible. In particular, system requires to identify the most relevant/significant contents of the text, extract them, order them, and return them to the user. Although extractive summarization task has been a popular research topic since 1958 (Luhn, 1958), yet it is a great challenge to summarize a text automatically using a computational system like a human generated summary. Several aspects about a good summary have been introduced by researchers. Das and Martins (2007) have discussed three major aspects for automatic text summarization.

- Summaries may be produced from a single or multiple documents.
- Summaries should consist of important information.
- Summaries should be concise.

These aspects are undoubtedly important, but a good summary should also consist of other aspects such as coverage, non-redundancy, cohesion, relevancy, and readability (Shareghi and Hassanabadi, 2008; Sankar and Sobha, 2009; Parveen et al., 2016). To incorporate all these aspects in a summary is a challenging task. This motivates us to improve the generated summaries in all these aspects.

II. Background

A. Text summarization phases

The automatic text summarization (ATS) is a process of finding a subset of document that contains the information residing in the entire document. According to Mani (1999), a text summarization system filters the significant information from the original document to generate an abbreviated version. Generally, the summarization process can be decomposed into three phases:

- Analysis of document text to obtain text representation.
- Transformation of text representation into summary representation.
- Transfiguration of summary representation into summary text to generate summary

The basic processing, elements, and resources, which are required to accomplish these phases are as follows.

1) Pre-processing: A high performance in the summarization system requires an effective pre-processing of the input text to obtain text representation. We accomplish this task of processing by employing Natural Language Tool Kit (NLTK). Here, the following steps are considered to preprocess the text.

- **Sentence Seperation:** It is a process of recognizing the individual sentences in a document which is used as a separate unit in summarization.
- **Stop words removal:** The process of stop-words removal eliminates the most frequent words occurring in a document like articles, prepositions, conjunctions, interrogations, helping verbs, etc. The stop words are removed due

*DOI:*http://dx.doi.org/10.37398/JSR.2020.640148

to their insignificant contribution in sentence extraction process.

- **Stemming:** It is a process of converting the semantically derived term into its morpheme term. We use the Porter stemmer for English text. According to Toman et al. (2006), the stemming process may draw a negative or insignificant impact in the performance of systems related to semantic analysis. So, we have experimented the proposed technique with both types of pre-processing (with/without stemming).
- **Part-of-Speech Tagging:** It is a process of identifying the part-of-speech words such as noun, adverb, verb, etc., in a sentence. However, the computational applications generally use more fine-grained POS tags like 'noun-plural'. Here, we have used the Stanford Log-linear POS tagger.
- **Keywords extraction:** In this step, we extracte the keywords from a document. Here, all the words other than stop words are considered as keywords.

2) Assessment of textual units: The major concept which has been used in transforming the document into summary representation is text features that can be exploited to find the relevant sentences of the document. In this paper, several features are used to score the sentences such as Aggregate similarity, Bushy path, Cue phrases, Lexical relation, Named entities, Noun and verb phrases, Numerical data, Open relations, Proper noun, Sentence centrality, Sentence length, Sentence position, Sentence with title words, Sentence significance, Frequent words (Verma and Om, 2016a,b,c, 2018, 2019a,b,c,d; Verma et al., 2019).

B. Evaluation approaches

Evaluations are done in three stages: co-selection based evaluation (with reference summary), content based evaluation (without reference summary), and document based evaluation (with original document), which are briefly described as follows.

1) Co-selection based Evaluations: The co-selection based evaluation relies on the co-occurrence of terms in system summary and it requires reference summary of documents for comparison. The evaluation is done by selecting the common terms of the system summary and reference summary. The related parameters for co-selection based evaluation are recall, precision, and F-score, as given below.

i. Recall: It is the ratio of total retrieved correct sentences to the total number of the retrieved correct sentences and non-retrieved correct sentences of a document. It can be estimated as follows.

$$Recall = \frac{\sum_{s \in sys} \sum_{gram_N \in s} Count_{match}(gram_N)}{\sum_{s \in ref} \sum_{gram_N \in s} Count(gram_N)} \quad (1)$$

Here, ref refers to reference summary, s stands for sentence, $Count_{match}(gram_N)$ is the maximum number of N-grams co-occurring in system summary and reference summary. $Count(gram_N)$ is the number of *N-gram* in reference summary.

ii. Precision: It is the ratio of total retrieved correct sentences to the total number of retrieved correct sentences and retrieved incorrect sentences of the document. It can be computed as follows.

$$Precision = \frac{\sum_{s \in sys} \sum_{gram_N \in s} Count_{match}(gram_N)}{\sum_{s \in sys} \sum_{gram_N \in s} Count(gram_N)} \quad (2)$$

Here, *sys* belongs to system summary and $Count(gram_N)$ is the number of *N-gram* in system summary.

iii. F-score: It measures the effectiveness of retrieval with respect to a user, which attaches β times as much importance to the recall as that of precision. The F-score for non-negative real β $(0 \leq \beta < \infty)$ is computed as follows.

$$F_\beta = \frac{(1+\beta^2)(Precision * Recall)}{(\beta^2 * Precision + Recall)} \quad (3)$$

iv. Improved Rates: We have also calculated the improved rates (*IR*) in the performance of the proposed methods with respect to other methods on the basis of above discussed parameters as follows.

$$IR = \frac{(PM - OM)}{OM} \quad (4)$$

where, *PM* is proposed method, *OM* is other method, and *IR* is improved rates.

2) Content based Evaluations: A co-selection method evaluates a summarization system on common terms. It cannot obtain connectivity of ideas, flow of sentences, relatedness of sentences with their previous sentences, and non-redundancy of contents in a summary. The content based method can address all these issues. We describe some content based evaluation methods that take into account different properties of a text. The content based evaluation only requires system summary. The related metrics for content based evaluation are as follows.

i. Cohesion: It is an essential parameter to capture the relations between concepts in a text. Halliday and Hasan (2014) identified five general categories of cohesive relations that are conjunction, reference, ellipsis, substitution, and lexical (Halliday and Hasan, 2014). The conjunction 'and' is the most basic and least cohesive relation between the clauses and the referential relation can be either anaphoric or cataphoric. An anaphoric relation occurs when a sentence refers back to something which has been previously explained, while cataphoric is just opposite to the anaphoric relation. The ellipsis relation occurs when, after a more specific explanation, the words are omitted in repeated phrase. In substitution relation, the words are not omitted as in ellipsis, but they are substituted by more general words instead of repeating words.

ii. Non-redundancy: The non-redundancy refers to the novelty in a summary. Several researchers have discussed how to obtain the non-redundant summary (Goldstein and Carbonell, 1998; Oufaida et al., 2014; Goldstein et al., 2000). They suggest that the summary should be non-redundant to increase the coverage of information residing in a document. So, it would be interesting to calculate the non-redundancy in the generated summary.

iii. Readability: The readability is also an essential parameter for measuring the performance of a summarizer. It tells about how easily a text content can be read and understood. The readability of a text can be measured in two aspects:

content and relatedness of a sentence with its previous sentence. The readability with reference to content depends on the complexity of vocabulary and syntax; whereas, the relatedness with previous sentence shows the fluency of reading.

In this paper, these metrics have been analysed manually by three expert assessors of English language. They are asked to give rating for each system generated summary on three levels likert scale format that are Yes, Partial, and No. The guidelines for readability based on cohesion are given as follows (DuBay, 2004).

1) Summary should consist of referential relation between the sentences wherever required.
2) Sentences of a summary should consist of ellipsis relation wherever required.
3) Summary should consist of substitution relation wherever required.
4) Summary should consist of lexical relation between the sentences.

They were instructed that if a summary follows all these guidelines, then it is rated as readable. If the summary follows half of these guidelines, then it is rated as partial readable; otherwise, it is rated as non-readable.

Human judgment cannot be consistent every time. So, it is interesting to measure how well two different judges agree on readability. The best way to measure for inter-judge agreement is the *kappa statistics* (Salton and McGill, 1986), which is defined as follows.

$$Kappa(\kappa) = \frac{P(A) - P(E)}{1 - P(E)} \tag{5}$$

where $P(A)$ is the proportion of the times the judges agreed, and $P(E)$ is the proportion of the times the judges agree by chance. The value of κ in the interval [2/3, 1] are seen as acceptable.

3) Statistical Testing: We perform the statistical test to determine whether the performances of the two methods are statistically significant or not. Thus, generally t-test (which is used when the sample size is small and two groups of the samples have been considered) has been applied on the datasets.

III. Challenges

In this section, several challenges are identified during summarizing the documents in the extractive manner, which are given as follows.

A. Problem of redundancy

Redundancy in asummary always has detrimental consequences on summarizing a document. A summary is more informative as much as it contains non-redundant contents. Most of the existing approaches focus on finding relevant content from document(s) and extract them to generate the summary. But, if we investigate about the redundancy, we can cover more information in the summary. In particular, similarity measurement plays a major role in finding the redundant contents in a document. If we can precisely measure the similarity between the contents of a document, then the redundancy can be minimized in the summary.

B. Problem of irrelevancy

The main aim of a summarization system is to extract relevant contents from a document that gives a quick view of the whole document. Generally, Human engineered text features are used to assess the sentences or textual units of a document. Since, it is not always feasible to incorporate all the considered features in a summary, some features may tend to create irrelevant contents in the summary. Thus, to consider all possible text features for assessment of the sentences increases complexity as well as irrelevancy. Hence, it is crucial to know which features are accountable for creating high quality summary in the given data. Moreover, in a reference summary, all the considered features do not manifest in the same ratio. Therefore, if we consider all the features in same ratio and assess the sentences accordingly, then this hypothesis may create irrelevant contents in the generated summary. Hence, it is also required to know proper ratio from the given dataset in which they should be presented in the summary.

C. Problem of loss of coverage

Coverage of topics of a document in the summary is an important aspect for generic text summarization. A good generic summary always reflects the information about every aspect mentioned in the document. However, it is not always necessary in the case of query based summarization. The current summarization techniques do not focus much on coverage of topics in the generated summaries. Hence, they fail to produce good summary in case of generic summarization. This problem arises mainly in the case of multi-document summarization where the number of topics in documents are much higher than in a single document.

In literature of text summarization, there exist some approaches which focus on maximizing the coverage while minimizing the redundancy. But these approaches do not gurantee maximization of coverage. Suppose they get best result at a point where the redundancy is minimum, then there is very high chances of loss of coverage.

D. Problem of non-readability and less cohesive content

A good summary should be readable and cohesive. By readable and cohesive mean that the contents of the summary should be conceptually related to each other. This paper presents a summarization method which takes into account readability and cohesion parameters to generate the summaries of the document. From the point of evaluation of the system generated summaries, this paper also presents a way to evaluate the summaries for these features in the summaries.

IV. Taxonomy of Summarization Techniques

There have been discussed a good number of works related to extractive text summarization as discussed below.

A. Graph based methods

In these methods, every sentence of a document is represented as a node of the graph and the relation between the sentences are denoted as edge. Every node is scored based

on the structured and non-structured text features, and the similarity between the sentences helps in traversing the graph in a significant manner. The extraction based multi-document summarization (Sripada et al., 2005) employs the WordNet based semantic similarity to find the similarity scores and to compute the sentence importance using ten text features. The unified approach for multi-document summarization relies on four assumptions to find the locally and globally important sentences in the documents through the affinity matrix of the sentence similarity (Wan, 2010). The G-FLOW (Christensen et al., 2013) is a system for coherent extractive multi-document summarization that generates an ordered summary by optimizing the coherence and salient factors, where the coherence of text is evaluated by an approximated discourse graph. Glavaš and Šnajder (2014) introduce an event based summarization method that exploits the strength of machine learning rule based approaches and performs effectively on the event oriented document collection.

The complex networks are also based on the graph theory. Amancio (2015) explores the linguistic properties for short written texts, which may be very helpful in summarization task. The topological properties of complex networks in short texts are investigated and it has been found that these can improve the global characterization of long texts also. Amancio et al. (2012) discuss a summarization method based on complex networks and syntactic dependencies. They employ some new metrics such as betweenness, vulnerability, closeness, and diversity, to extract the sentences from documents and they find that the diversity metric is the best for extraction. It is further suggested that the syntactic parsing can enhance the performance of summarizers. Tohalino and Amancio (2018) discuss a multilayer approach based extractive summarizer where several measurements such as degree, strength, page rank, accessibility, symmetry, shortest path, absorption time, etc., are used to weight the edges in the network of documents. They find that the distinction between intra- and inter-layer edges can play a major role in improving the results of a summarizer. Liu et al. (2018) present a survey on the graph based summarization methods by categorizing the state-of-the-art methods according to the input graphs and associated approaches. This survey leads to several open research areas such as temporal graph summarization for document, improvements in standardizing, generalizing algorithms, etc. The general limitation with these methods is that these are poorly applicable to large scale of data. So, their performance may be limited to single document summarization. To address this limitation, we propose the summarization methods based on meta-heuristic approaches, which can be efficiently applied to the large size of documents.

B. Maximal Marginal Relevance based methods

In maximum marginal relevance (MMR) based methods, the summarization task is modeled in such a way that the contents of produced summary should consist of relevant information to the query as well as minimal similarity among the contents. Goldstein and Carbonell (1998) discuss a maximal marginal relevance based method for multi-document text summarization. This method balances the coverage and relevancy with query factors in the summary. The balancing weight for relevancy is taken as 0.7, and 0.3 for non-redundancy. Goldstein et al. (2000) focus on the issues in multi-document summarization that are compression, speed, redundancy, and passage selection. Wang et al. (2009) discuss a summarization method for e-mail data by analyzing the relationship between the MMR model and content cohesion in e-mails that enhances the precision scores. Chaudhari and Mattukoyya (2018) present an MMR model with Naive-based tone biasing model. The content generated from MMR model is used as an input for the Naive biasing model using a set of polarity tags. Some of the MMR based methods are discussed in meta-heuristic based methods that consider them as an optimization problem. The general limitation with these methods is that their performances do not guarantee for the presence of both coverage and non-redundancy aspects in the summary. To address this limitation, we propose clustering based summarization methods that address both aspects.

C. Meta-heuristic based methods

For last couple of years, many researchers have focused on the optimization algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Harmony Search (HS), Differential Evolution (DE), and Cat Swarm Optimization (CSO) for single as well as multi-document summarization. To our best knowledge, the genetic algorithm was used first time in the document summarization task in (He et al., 2006) to retrieve the relevant sentences based on four summary factors: satisfied length, high coverage, high informativeness, and low redundancy. This method takes into account the similarity between the words using the WordNet that is further used to find the term frequency. Kogilavani and Balasubramanie (2010) present a method for optimal summary generation by grouping the related documents into a cluster and extracts the important sentences from each cluster using the genetic algorithm. The multi-criteria optimization based multi-document summarization (John et al., 2017) finds the extractive generic summary with maximal relevance and minimal redundancy. The sentences are scored using five features: TF-IDF, aggregate cross sentence similarity, title similarity, proper noun, and sentence length. Rautray and Balabantaray (2017) discuss a technique for summarization using the cuckoo search using three features: coverage, cohesion, and readability for summarization. Alguliev et al. (2013) discuss a summarization technique based on differential evolution that focuses on three aspects of summarization: content coverage, diversity, and length of summary; and optimizes these aspects using differential evolution. The cosine function is used for similarity measurement. A PSO based text summarization model (Alguliev et al., 2011) focuses on maximizing the content coverage and minimizes the redundancy in the summary. This method optimizes the relevancy, redundancy, and summary length together through a binary PSO and the Google hit based dissimilarity function (NGD). The limitation of these methods is that the metaheuristic approaches used for the summarization generally get trapped into local optima. Moreover, these techniques do

not provide information about the behavior of a function such as steepness and extrema in the search space. Thus, a gradient based optimization approach is used in the proposed method so that the convergence of the search algorithm can be drastically enhanced.

D. Other methods

In recent years, some summarization methods have been discussed based on Reinforcement Learning (RL). The main objective of RL is to facilitate the optimization of non-differentiable functions such as ROUGE. In this respect, Narayan et al. (2018) discuss an extractive summarization which is globally trained by optimizing the ROUGE function. It consists of three main modules: sentence encoder (implemented with Convolutional Neural Networks (CNNs) for continuous representation of sentences), document encoder (implemented with Recurrent Neural Networks (RNNs) + Long Short-Term Memory (LSTM) to avoid the vanishing gradient problem), and sentence extractor (implemented with RNNs + LSTM for summarization). It also uses the combination of maximum-likelihood cross-entropy loss and rewards obtained by the policy gradient RL as the objective function. Paulus et al. (2017) discuss an RL+ML based model for abstractive summarization where a key attention mechanism and learning objective to address the redundancy problem has been introduced. Here, the sentence encoder and decoder are based on RNNs. Experimentally, this model is effective for long document summarization that suggests ROUGE should not be the only metric to optimize in summarization. Lee and Lee (2017) discuss a deep Q-Network based single document summarization model that uses both content and position based embedding features to select the sentences for summary.

Here, we present some state-of-the-art summarization methods that are based fuzzy, evolutionary, and clustering algorithms and their hybrid form. Aretoulaki (1997) presents a model for abstract generation that uses four processes: symbolic morphological, syntactic, semantic, and pragmatic processes. Each process is responsible for collecting the specific feature based information that is passed to an artificial neural network (ANN) to score the textual units. Thione et al. (2004) discuss a model that exploits the strengths of summarist and PULSUMM base tree-based summarization systems. Chang et al. (2008) discuss a method that combines K-mixture term weighting method and linguistic method to generate the summary. Alonso i Alemany and Fuentes Fort (2003) discuss a summarization model that combines the cohesive properties of the text with coherence relations and strengthens the lexical chain based summarizer using the rhetorical and argumentative structures. Da Cunha et al. (2007) present a hybrid summarization model for Spanish text that integrates the Cortex (inspired by linguistic technique) as well as Enertex (inspired by statistical physics) for summarization. Binwahlan et al. (2010) discuss two hybrid models by exploiting the strength of fuzzy logic, evolutionary algorithm, and maximal marginal importance algorithm, and evaluate them on the DUC02 dataset. Abbasi-ghalehtaki et al. (2016) discuss a summarization model based on fuzzy logic, evolutionary algorithms, and cellular learning automata (CLA) in which the similarity function is modeled as a combination of CLA and artificial bee colony (ABC) problem. The weights of text features are customized by using the PSO and GA and it is compared with fourteen other summarization methods. Mehta and Majumder (2018) present a comparative study of various existing text summarizers with respect to the sentence ranking, sentence similarity, and text representation. They suggest that the combination of different techniques (or a hybrid model) of summarization in a systematic manner can enhance the performance of the summarization system. Goularte et al. (2019) discuss a method for text summarization using the fuzzy rules which is further used for automatic text assessment. They show that the fuzzy based summarization system can successfully enhance the quality of the generated summaries. Hu et al. (2017) discuss a clustering based summarization method for opinion data (Opinosis) where they suggest that the clustering of reviews can play a major role in coverage of reviews related to every instance. Wang et al. (2017) discuss nine heuristic based methods for sentence extraction from long documents. They suggest that the removal of redundant contents from a document can speed up the ability of summarization system for finding relevant sentences and summary. Tayal et al. (2017) discuss a summarization method based on soft computing techniques that cluster the sentences of a document for finding similar sentences and merge them according to their similarities. He et al. (2016) discuss a multi-document summarization method based on group sparse learning which acquires structural knowledge among the group of sentences. They use the Nesterov's method to optimize the group sparse convex for better convergence behavior. Wei et al. (2016) present a summarization method, called Heterogeneous Feature Symmetric Summarization (HFSS). Abdi et al. (2018) present a query based summarization method, called QMOS, which is a two-stage procedure: sentiment analysis and summarization. The semantic sentiment analysis is carried out by combining multiple sentiment lexicons. The summarization is done on the basis of syntactic and semantic analysis of sentences. Azmi and Altmami (2018) discuss an abstractive summarization approach for Arabic text document, which uses a sentence reduction approach and rhetorical structural theory based sentence extraction approach to generate summary. Mosa et al. (2018) present a survey on swarm intelligence (SI) based summarization techniques and report that the usage of SI approaches is quite limited with respect to summarization task. They discuss a summarization framework to cover multi-objective optimization task using SI. Sanchez-Gomez et al. (2018) discuss a multi-objective ABC based summarization method and its efficacy is shown on the DUC02 dataset.

V. Conclusion

In this paper, we have presented a technical background for document summarization. This paper has also discussed several challenges as well as surveys of the existing summarization methods. From these discussions, we have observed that many techniques suffer from various challenges, for example, the graph based methods have imitation in data size,

the clustering based methods require prior knowledge of the number of clusters, the MMR approaches have uncertainty for the coverage and non-redundancy aspects in the summary, etc. So, it is imperative that further research is required in this field to develop more effective methods for document summarization.

References

Abbasi-ghalehtaki, R., Khotanlou, H., and Esmaeilpour, M. (2016). Fuzzy evolutionary cellular learning automata model for text summarization. *Swarm and Evolutionary Computation*, 30:11–26.

Abdi, A., Shamsuddin, S. M., and Aliguliyev, R. M. (2018). Qmos: Query-based multi-documents opinion-oriented summarization. *Information Processing & Management*, 54(2):318–338.

Alguliev, R. M., Aliguliyev, R. M., Hajirahimova, M. S., and Mehdiyev, C. A. (2011). Mcmr: Maximum coverage and minimum redundant text summarization model. *Expert Systems with Applications*, 38(12):14514–14522.

Alguliev, R. M., Aliguliyev, R. M., and Isazade, N. R. (2013). Multiple documents summarization based on evolutionary optimization algorithm. *Expert Systems with Applications*, 40(5):1675–1689.

Alonso i Alemany, L. and Fuentes Fort, M. (2003). Integrating cohesion and coherence for automatic summarization. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 1–8. Association for Computational Linguistics.

Amancio, D. R. (2015). Probing the topological properties of complex networks modeling short written texts. *PloS one*, 10(2):e0118394.

Amancio, D. R., Nunes, M. G., Oliveira Jr, O. N., and Costa, L. d. F. (2012). Extractive summarization using complex networks and syntactic dependency. *Physica A: Statistical Mechanics and its Applications*, 391(4):1855–1864.

Aretoulaki, M. (1997). Towards a hybrid abstract generation system. In *New Methods in Language Processing*, pages 55–68. UCL Press.

Azmi, A. M. and Altmami, N. I. (2018). An abstractive arabic text summarizer with user controlled granularity. *Information Processing & Management*, 54(6):903–921.

Binwahlan, M. S., Salim, N., and Suanmali, L. (2010). Fuzzy swarm diversity hybrid model for text summarization. *Information processing & management*, 46(5):571–588.

Chang, T.-M., Hsiao, W.-F., et al. (2008). A hybrid approach to automatic text summarization. In *2008 8th IEEE International Conference on Computer and Information Technology*, pages 65–70. IEEE.

Chaudhari, M. and Mattukoyya, A. N. (2018). Tone biased mmr text summarization. *arXiv preprint arXiv:1802.09426*.

Christensen, J., Soderland, S., Etzioni, O., et al. (2013). Towards coherent multi-document summarization. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1173.

Da Cunha, I., Fernández, S., Morales, P. V., Vivaldi, J., SanJuan, E., and Torres-Moreno, J. M. (2007). A new hybrid summarizer based on vector space model, statistical physics and linguistics. In *Mexican International Conference on Artificial Intelligence*, pages 872–882. Springer.

Das, D. and Martins, A. F. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4(192-195):57.

DuBay, W. H. (2004). The principles of readability. *Online Submission*.

Glavaš, G. and Šnajder, J. (2014). Event graphs for information retrieval and multi-document summarization. *Expert systems with applications*, 41(15):6904–6916.

Goldstein, J. and Carbonell, J. (1998). Summarization:(1) using mmr for diversity-based reranking and (2) evaluating summaries. In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*, pages 181–195. Association for Computational Linguistics.

Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pages 40–48. Association for Computational Linguistics.

Goularte, F. B., Nassar, S. M., Fileto, R., and Saggion, H. (2019). A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert Systems with Applications*, 115:264–275.

Halliday, M. A. K. and Hasan, R. (2014). *Cohesion in english*. Routledge.

He, R., Tang, J., Gong, P., Hu, Q., and Wang, B. (2016). Multi-document summarization via group sparse learning. *Information Sciences*, 349:12–24.

He, Y.-X., Liu, D.-X., Ji, D.-H., Yang, H., and Teng, C. (2006). Msbga: A multi-document summarization system based on genetic algorithm. In *Machine Learning and Cybernetics, 2006 International Conference on*, pages 2659–2664. IEEE.

Hu, Y.-H., Chen, Y.-L., and Chou, H.-L. (2017). Opinion mining from online hotel reviews–a text summarization approach. *Information Processing & Management*, 53(2):436–449.

John, A., Premjith, P., and Wilscy, M. (2017). Extractive multi-document summarization using population-based multicriteria optimization. *Expert Systems with Applications*, 86:385–397.

Kogilavani, A. and Balasubramanie, P. (2010). Clustering based optimal summary generation using genetic algorithm. In *Communication and Computational Intelligence (INCOCCI), 2010 International Conference on*, pages 324–329. IEEE.

Lee, G. H. and Lee, K. J. (2017). Automatic text summarization using reinforcement learning with embedding features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 193–197.

Liu, Y., Safavi, T., Dighe, A., and Koutra, D. (2018). Graph summarization methods and applications: A survey. *ACM Computing Surveys (CSUR)*, 51(3):62.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Mani, I. (1999). *Advances in automatic text summarization*. MIT press.

Mehta, P. and Majumder, P. (2018). Effective aggregation of various summarization techniques. *Information Processing & Management*, 54(2):145–158.

Mosa, M. A., Anwar, A. S., and Hamouda, A. (2018). A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms. *Knowledge-Based Systems*.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Oufaida, H., Nouali, O., and 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.

Parveen, D., Mesgar, M., and Strube, M. (2016). Generating coherent summaries of scientific articles using coherence patterns. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 772–783.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Rautray, R. and Balabantaray, R. C. (2017). An evolutionary framework for multi document summarization using cuckoo search approach: Mdscsa. *Applied Computing and Informatics*.

Salton, G. and McGill, M. J. (1986). Introduction to modern information retrieval.

Sanchez-Gomez, J. M., Vega-Rodríguez, M. A., and Pérez, C. J. (2018). Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. *Knowledge-Based Systems*, 159:1–8.

Sankar, K. and Sobha, L. (2009). An approach to text summarization. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, pages 53–60. Association for Computational Linguistics.

Shareghi, E. and Hassanabadi, L. S. (2008). Text summarization with harmony search algorithm-based sentence extraction. In *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, pages 226–231. ACM.

Sripada, S., Kasturi, V. G., and Parai, G. K. (2005). Multi-document extraction based summarization. *CS 224N, Final Project*.

Tayal, M. A., Raghuwanshi, M. M., and Malik, L. G. (2017). Atssc: Development of an approach based on soft computing for text summarization. *Computer Speech & Language*, 41:214–235.

Thione, G. L., Van den Berg, M., Polanyi, L., and Culy, C. (2004). Hybrid text summarization: Combining external relevance measures with structural analysis. *Text Summarization Branches Out*, pages 51–55.

Tohalino, J. V. and Amancio, D. R. (2018). Extractive multi-document summarization using multilayer networks. *Physica A: Statistical Mechanics and its Applications*, 503:526–539.

Toman, M., Tesar, R., and Jezek, K. (2006). Influence of word normalization on text classification. *Proceedings of InSciT*, 4:354–358.

Verma, P. and Om, H. (2016a). Extraction based text summarization methods on users review data: A comparative study. In *International Conference on Smart Trends for Information Technology and Computer Communications*, pages 346–354. Springer.

Verma, P. and Om, H. (2016b). A survey on indian language text summarization techniques, evaluation, and existing tools. In *International conference on Innovative Systems*, page 32.

Verma, P. and Om, H. (2016c). Theme driven text summarization using k-means with gap statistics for hindi documents. In *International Conference on Computing, Communication and Sensor Network*, pages 90–94.

Verma, P. and Om, H. (2018). Fuzzy evolutionary self-rule generation and text summarization. In *15th International Conference on Natural Language Processing*, page 115.

Verma, P. and Om, H. (2019a). Collaborative ranking-based text summarization using a metaheuristic approach. In *Emerging Technologies in Data Mining and Information Security*, pages 417–426. Springer.

Verma, P. and Om, H. (2019b). Mcrmr: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. *Expert Systems with Applications*, 120:43–56.

Verma, P. and Om, H. (2019c). A novel approach for text summarization using optimal combination of sentence scoring methods. *Sādhanā*, 44(5):110.

Verma, P. and Om, H. (2019d). A variable dimension optimization approach for text summarization. In *Harmony Search and Nature Inspired Optimization Algorithms*, pages 687–696. Springer.

Verma, P., Pal, S., and Om, H. (2019). A comparative analysis on hindi and english extractive text summarization. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):30.

Wan, X. (2010). Towards a unified approach to simultaneous single-document and multi-document summarizations. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1137–1145. Association for Computational Linguistics.

Wang, B., Liu, B., Sun, C., Wang, X., and Li, B. (2009). Adaptive maximum marginal relevance based multi-email summarization. In *International Conference on Artificial Intelligence and Computational Intelligence*, pages 417–424. Springer.

Wang, W., Li, Z., Wang, J., and Zheng, Z. (2017). How far we can go with extractive text summarization? heuristic methods to obtain near upper bounds. *Expert Systems with Applications*, 90:439–463.

Wei, W., Ming, Z., Nie, L., Li, G., Li, J., Zhu, F., Shang, T., and Luo, C. (2016). Exploring heterogeneous features for query-focused summarization of categorized community answers. *Information Sciences*, 330:403–423.