Extractive and Abstractive Text Summarization Techniques

PL.Prabha, M.Parvathy

Abstract: ***Text summarization generates an abstract version of information on a particular topic from various sources without modifying its originality. It is essential to dig information from the large repository of data, thereby eliminating the irrelevant information. The manual summarization consumes a large amount of time and hence an automated text summarization model is required. The summarization can be performed from a single source or multiple sources. The Natural Language Processing (NLP) based text summarization can be generally categorized as abstractive and extractive methods. The extractive methods mine the essential text from the document whereas the abstractive methods summarize the document by rewriting. The extractive summarization methods rely on topics and centrality of the document. The abstractive techniques transform the sentences based on the language resources available. This paper deals with the study of extractive as well as abstractive strategies in text summarization. Overall objective of this paper is to provide a significant direction to the researchers to learn about different strategies applied in text summarization.***

Index Terms: Text summarization, Abstractive Summarization, Extractive Summarization, Natural Language Processing.

I. INTRODUCTION

The enormous growth of internet had led to the overwhelming content in the form of web pages that generates information on daily basis. The content of internet may be in the form of news, images, email, database, videos and so on [1]. The collection of relevant documents has become a tedious task due to the large amount of information available over the web. Hence, Natural Language Processing (NLP) is being utilized to scrap the relevant documents from the scattered resources. Text summarization is a sub branch of NLP which supports the user to refine the necessary points from a large document [2][3]. This summarization can also be applied to extract important summary from multiple documents. It facilitates the user to reduce the time of reading multiple sources thereby fastening the searching process. A good summarization should produce the key points from the input in a concise and fluent format [4]. Text summarization find its application in Newspaper articles,Stock Market,Scientific paper,Weather forecast. The process of text summarization consists of source identification, content interpretation and summary generation. In identification, the important points are identified using segmentation and stemming for summary generation. The identified points are modified during the content interpretation, which involves ranking the phrase and assigning weights. The entire summary is reformulated using the key words, stop words, and removing redundant phrases to make the content readable and understandable in the generation phase [5]. The general flow of text summarization process is shown in Fig 1.

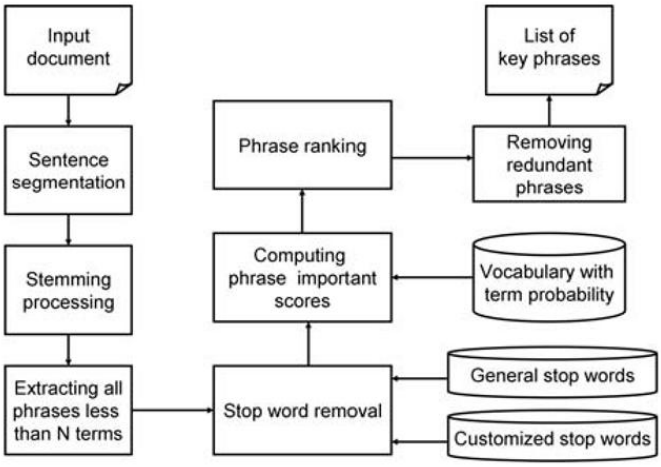

Figure 1: Process of Text Summarization

In general, text summarization is categorized as extractive and abstractive summarization. In extractive method the important content is collected from different sources and combined into a new document whereas, in abstractive summarization the new document is created by paraphrasing the sentences. The extractive summarization can be further classified as topic based extraction and centrality based extraction [5]. Each and every word in the document plays a vital role in topic based methods. The words are maintained in a corpus or a dictionary to assign the weights relating to the importance. The centrality based strategies work based on the rank assigned to each sentences with respect to centroid. Both the extractive and abstractive has its pros and cons. The summary of the extractive methods are tedious to understand because of the sentence complexity.

The extraction may lead to irrelevance of words placed together in a sentence which are highly conflict. In case of multi document summarization, it results in context mismatch, inaccuracy and low fidelity [6]. In abstractive summarization, the problem of representation is a major challenge which leads to an unstructured document.

Revised Manuscript Received on April 27, 2020.
***** Correspondence Author
PL.Prabha, Research Scholar,CSE Department, K.L.N. College of Information Technology, Pottapalyam,Sivagangai Dist.
Dr.M.Parvathy, Professor , CSE Department, Sethu Institute of Technology, Pulloor,Kariapatti

Hence, evaluation plays a vital role in assessing the quality of the summarized document. The evaluation methods can either done by manual evaluation techniques or retrieval techniques [7]. The summarization process considers several features including words in the title of the document, location of sentence, length of the sentence, lower and upper case, and sentence phrases. In addition, cohesion is also required to find out the similarity between sentences and between centroids and sentences. A good summary should avoid the words that are not essential to the document as well as it should have a check on redundancy.

The paper is organized as follows: Section II reviews the state-of-art techniques of extractive text summarization in single as well as multiple documents. Section III has a brief review of abstractive text summarization models. Section IV has a detailed discussion of the reviewed techniques based the techniques used in each strategy. Section V concludes the review with future insights

II. EXTRACTIVE SUMMARIZATION MODELS

Extractive methods utilize conventional methods including word frequency, cluster of words, graph based approaches, machine learning and neural network based approaches.

A. Word Frequency based Summarization

The word contained in different sentences of the document is identified to weigh them according to their frequency. The average count of a particular word available in a document is termed as word frequency. The search query and information retrieval process is carried out in frequency calculation. The sentences that have high scores are extracted to frame the summary based on the similarity score of the sentence and the submitted query. This method can be adopted both for query based as well as generic summarization [8].

B. Cluster based Extractive Summarization

In cluster based extraction, the summary of the document is clustered based on the various topics available in the document. In general, a document consists of several sections on different contexts. The precise summary can be generated out of entirely different contexts organized to form a document. The concept of word frequency is also incorporated in the cluster based approach. In some cases, the input to the summarizer model is the clustered document, which is then ranked according to the word frequency. The sentences are chosen with respect to the similarity and its location [9].

C. Graph based Summarization

Initially, once after identifying different topics in the document the stop words and stem words are identified and placed as nodes of the undirected graph [10]. An edge is drawn between two different nodes as in the following cases:

- If similarity index between the sentences exceeds the threshold value.
- If there is common words between the sentences.

The joining edges help the summarizer to easily map the related sentences belonging to a particular a topic. The number of edges connected to a particular node is termed as cardinality. The highest the cardinality of a node, the most important the sentences are.

D. Machine Learning Techniques for Summarization

The text summarization techniques include statistical based method, machine learning and Artificial Intelligence based techniques. In statistical based text summarization, the meaning of the words is not considered for summarization instead it uses the physical features. The statistical methods apply cue phrases, word frequency, title words and location of the sentence [11]. Even though, the statistical methods are easy to implement, the summarization quality is low. Hence, in recent time, the use of statistical methods for text summarization is not recommended. The machine learning based text summarization techniques utilizes training dataset to train the model, which is later evaluated by the test data. The sentences are selected based on the probability of the suitability of sentences. It utilizes several classification algorithms such as Naïve Bayes, fuzzy logic and neural networks. The probability of sentences that can be placed in a meaningful manner is predicted using Bayes theorem of probability [12]. The machine learning based methods results in good summarization quality as it considers several features including title, sentence location, thematic words, sentence length and number of words in the title.

E. Summarization based on Neural Networks

The feed forward neural network can be used with three layers for training the NLP model using the sentences to be added in the summary. Once the training phase is completed, the model learns to identify the relevant and irrelevant sentences which can be tested using test documents. The feature fusion phase is followed by the training phase that involves shuffling the relevant features and removal of irrelevant features. The weight of connections is computed, so that the connections with less weight can be eliminated for performance enhancement. The irrelevant features are identified through weights. In the hidden layer, the adaptive clustering technique is applied to cluster the relevant features based on the frequency and centroid. The relevant features are shuffled by replacing the activation value with the centroid value. These two steps help in ranking the sentences through with the output can be a meaningful and precise summary [13].

F. Fuzzy based Summarization

The Fuzzy **based** system has a knowledge base that provides rules and summarization is proceeded according to the length of the sentence, similarity of the keywords and sentences. According to the rules and the characteristics of the sentence, all the sentences are assigned with a value ranging from zero to one. Each feature has any either significant or non significant values among which high and very high are significant whereas very low, low and medium are insignificant in nature. A common fuzzy logic system has fuzzifier, defuzzifier, knowledge base and an inference engine. A membership function in the fuzzifier converts the crisp values into the above mentioned values. The inference engine utilizes the IF THEN rules available in the knowledge base fed by the expert. The sentence score is computed by the defuzzifier in which the values are again converted to crisp values [14][15].

G. Regression based Summarization

The weight of the features is computed using the mathematical regression model, where dependent values are generated based on the independent values given as input. The training phase establishes a relationship between the inputs and outputs using the regression equation. The performance of the model is evaluated in the testing phase [16].

H. Query based Summarization

The word frequency is identified and the sentences that are present in the search query are given a high score when compared to the other sentences. Such sentences are included in the summary in addition to the sentences derived from the other sections. The size of the frame is set in such a way that there should be no need for scrolling. The main titles and the sub titles are included in the summary. The sentences are included in the summary in accordance with the rank until the maximum frame size in reached. The query based summarization also utilizes the Bayesian method for focusing the importance sentences [17].

I. Multilinguistic Summarization

In multilingual summarization the sources are in multiple languages and the final summary obtained must be in a particular language. The unary and composite features are adopted to calculate the similarity among them. The new features can be added during run time which makes the model more dynamic. The document structure, syntactic and statistical features are considered to make the system robust and interactive. In some cases the documents are translated to XML format and several features are derived from each cluster. The dependencies between the sentences are identified by computing the composite score [18].

J. Summarization form Multiple Documents

The more informative and concise summaries can be generated through multi document summarization. The interesting parts of each document is gathered and combined to form the summary. The content from each source is selected and filtered before summarization. The unigrams, bigrams, and trigrams are identified to compute the likelihood ratio using which the sentences are ranked. The pronouns, verbs, and conjunctions are eliminated to make the document to get rid of discontinuity. The marginal relevancy filter can be applied to reduce redundancy, which is performed by calculating the overlap ratio [19].

III. ABSTRACTIVE SUMMARIZATION MODELS

In recent times the abstractive summarization is high used in NLP models as it forms new sentence to summarize the document. It generates the document in a simple language and error free manner, thereby enhancing the quality of summary [20]. The abstractive mechanism can be further sub divided as structure based and semantic based techniques. The structure based approach derives data from the templates, ontology, tree and graph based structures. The NLP models are adopted in semantic based approaches that categorize the nouns and verbs from the dataset.

A. Tree based Abstraction

The dependencies of the sentences in a document can be represented as trees using the language generators. The theme of the summary is identified from the multiple documents given as input to the system. The clustering and sentence fusion methods are used to combine the sentences one after the other which results in a statistical summary. The phrases that provide common information are fused to form the summary [21].

B. Ontology based Abstractive Summarization

Ontology refers to a knowledge base in which multiple documents regarding similar topics are available. The information in the ontology is structured by the domain experts. The meaningful sentences are generated from the dictionary or corpus created by the experts. The membership degrees are created using the ontology based fuzzy logic which results in a precise summary of a topic [22].

C. Abstractive Summarization based on Templates

The general overview of a theme is generated as text snippets which are already interlinked with the guiding slots. The rules and patterns for extraction are mapped to the text snippets. These templates can generate summaries based on either extraction or abstraction. The information are stored in the databases, from which the entire summary is updated [23].

D. Rule based Abstraction

The outline of the document is formed using the generation patterns thereby identifying the classes and choice of content. The nouns and verbs in the phrases are related in a semantic manner. The word graph is constructed using the rules in knowledge base which is used for combining the sentences. The dataset undergoes cross validation to spilt the training and testing data. The training set is classified either are suitable and non suitable sentence. The suitable sentence is included in the summary to improve the relevance and to reduce the redundancy [24].

E. Multimode Semantic Abstraction Model

The multimode model has both text and images in the summarized document. The large amount of symbolic and numerical data is summarized using the NLP models that interpret data into text format. The process of linearization and morphological operations are undergone by the input documents to identify the lexical items. The features chosen for training the model reflects the quality of summary [25].

F. Graph based Semantic Summarization

Rich semantic graphs are constructed from the input document which is further reduced to form the final abstract of the summary. A set of related words are generated based on ontology from each word in the input document. Several rich semantic sub graphs are generated using the concept validation process. The sub graphs are further consolidated by removing and merging the nodes based on heuristic rules. The ontology is formed based on the nouns and verbs involved in the document. Once the lexical analysis gets done, the document is translated to syntactic format. The nouns are categorized into three, namely, main noun, subject noun and object noun and those nouns are place in the nodes of the graph [10].

G. Informative Summarization

Despite of sentences from the input document, the abstracts are considered for further summarization. The characteristics and attributes of the entities in the documents are identified for selecting the contents. The output of the informative model does not produce sentences instead they produce information items which are then clubbed together to form the summary. The local decisions are ranked and sorted to form the global decisions [26].

H. Text Representation based Semantic Models

The predicate of the sentence is identified and ranked to prioritize them based on its significance using the NLP tool. Each sentence in the document is segregated and assigned a semantic score computed from the similarity matrix of semantic graph. The structure of the predicate is determined form the modified graph ranking and it results in a precise summary with minimum redundancy [24].

IV. DISCUSSION AND ANALYSIS

The significance of text summarization is to achieve easy reading of the document by generating shorter summaries without any content loss. The summary generated by abstractive methods is more beneficial than the extractive methods as it uses linguistic models to frame sentence on its own without extracting the same content from the document. In abstractive summarization, the structure based approaches are found to be highly significant as it is free from grammatical errors. The semantic relation affects the quality of the semantic based models thereby reducing the quality, but the semantic models provides high cohesion with minimum redundancy. As the frameworks of the abstractive models have no standardization, abstractive summarization is considered to be a bit difficult. The process of paraphrasing, lexical substitutions and sentence reformulation are the tedious tasks of abstractive summarization.

Table I :Extractive and Abstractive Text Summarization Technique

Reference No	Summarization Method	Technique	Proposed Model
[8]	Extractive	Term Frequency Machine Learning	Data preprocessed based on Term Frequency Clustering based classifier for Arabic text documents
[9]	Extractive	Clustering	COSUM Model (Clustering and Optimization based Summarization)
[10]	Extractive	Graph based Summarization	Modified Text Rank Summarization
[11]	Extractive	Statistical methods	Statistical features based summarization
[12]	Extractive	Machine Learning	Naive Bayes classification based summarization
[13]	Extractive	Unsupervised learning	Neural Networks based summarization
[14]	Extractive	Fuzzy Logic	Summarization on Brazilian Portuguese text
[16]	Extractive	Regression based Summarization	Regression for sentence prediction
[17]	Extractive	Query based summarization	Machine Reading Comprehension model
[18]	Extractive	Summarization from multiple languages	Ranking method
[19]	Extractive	Summarization from multiple documents	Fuzzy logic and ranking
[21]	Abstractive	Tree structure for summarization	Recurrent Neural Network and Syntax Tree
[22]	Abstractive	Ontology based summarization	Ontology and Genetic Algorithm
[25]	Abstractive	Summarization using Semantics	Deep learning
[26]	Abstractive	Summarization using collected information	Swarm Intelligence Algorithms

V. CONCLUSION

The importance of text summarization is discussed briefly with the significance of precise and consistent summaries that helps the reader to utilize their time efficiently. The overall objective of text summarizations is to eliminate the redundancy and to present a clear and concise summary without any change in the meaning. This paper has surveyed the existing techniques of text summarization in two broad aspects, namely, extractive and abstractive summarization. The different approaches in extractive and abstractive summarization have been studied and the strategies adopted in each approach in discussed. The paper concludes that the summaries generated by abstractive approaches are better than the extractive approaches.Evaluation Results of using DUC dataset ROUGE-1 Score for Tree based Abstraction is 0.3-0.4,Ontology based Abstraction is 0.29-0.319,Template is 0.21-0.5,Multimode Abstraction is 0.05 -0.3,Graph based is 0.31and Deep learning is 0.28 -0.47. Hence Abstractive summaries with Deep Learning produce content rich and cohesive summary . In future, there is a great insight of opportunities for the researches to adopt abstractive summarization techniques along with machine learning and deep learning.

REFERENCES

1. Gupta, Vanyaa, Neha Bansal, and Arun Sharma. "Text summarization for big data: A comprehensive survey." *International Conference on Innovative Computing and Communications*. Springer, Singapore, 2019.
2. Zhang, Haoyu, et al. "Pretraining-based natural language generation for text summarization." *arXiv preprint arXiv:1902.09243* (2019).
3. Kanapala, Ambedkar, Sukomal Pal, and Rajendra Pamula. "Text summarization from legal documents: a survey." *Artificial Intelligence Review* 51.3 (2019): 371-402.
4. Roul, Rajendra Kumar, et al. "A new automatic multi-document text summarization using topic modeling." *International conference on distributed computing and internet technology*. Springer, Cham, 2019.
5. Mao, Xiangke, et al. "Extractive summarization using supervised and unsupervised learning." *Expert Systems with Applications* 133 (2019): 173-181.
6. Xiao, Wen, and Giuseppe Carenini. "Extractive summarization of long documents by combining global and local context." *arXiv preprint arXiv:1909.08089* (2019).
7. Behal, Sonali, Aayush Gupta, and Vivek Kumar Sehgal. "Automatic Text Summarization using Natural Language Processing." (2019).
8. Sangaiah, Arun Kumar, et al. "Arabic text clustering using improved clustering algorithms with dimensionality reduction." *Cluster Computing* 22.2 (2019): 4535-4549.
9. Alguliyev, Rasim M., et al. "COSUM: Text summarization based on clustering and optimization." *Expert Systems* 36.1 (2019): e12340.
10. Mallick, Chirantana, et al. "Graph-based text summarization using modified TextRank." *Soft Computing in Data Analytics*. Springer, Singapore, 2019. 137-146.
11. Meena, Yogesh Kumar, and Dinesh Gopalani. "Statistical Features for Extractive Automatic Text Summarization." *Natural Language Processing: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2020. 619-637.
12. Shah, Chintan, and Anjali Jivani. "An Automatic Text Summarization on Naive Bayes Classifier Using Latent Semantic Analysis." *Data, Engineering and Applications*. Springer, Singapore, 2019. 171-180.
13. Alami, Nabil, Mohammed Meknassi, and Noureddine En-nahnahi. "Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning." *Expert systems with applications* 123 (2019): 195-211.
14. Goularte, Fábio Bif, et al. "A text summarization method based on fuzzy rules and applicable to automated assessment." *Expert Systems with Applications* 115 (2019): 264-275.
15. Kumar, A. K. S. H. I., and A. D. I. T. I. Sharma. "Systematic literature review of fuzzy logic based text summarization." *Iranian Journal of Fuzzy Systems* 16.5 (2019): 45-59.
16. Zopf, Markus, Eneldo Loza Mencía, and Johannes Fürnkranz. "Which Scores to Predict in Sentence Regression for Text Summarization?." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018.
17. Egonmwan, Elozino, Vittorio Castelli, and Md Arafat Sultan. "Cross-Task Knowledge Transfer for Query-Based Text Summarization." *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. 2019.
18. Wan, Xiaojun, et al. "Cross-language document summarization via extraction and ranking of multiple summaries." *Knowledge and Information Systems* 58.2 (2019): 481-499.
19. Patel, Darshna, Saurabh Shah, and Hitesh Chhinkaniwala. "Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique." Expert Systems with Applications 134 (2019): 167-177.
20. Gerani, Shima, Giuseppe Carenini, and Raymond T. Ng. "Modeling content and structure for abstractive review summarization." Computer Speech & Language 53 (2019): 302-331.
21. Zhang, Jian, et al. "A novel neural source code representation based on abstract syntax tree." 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). IEEE, 2019.
22. Sharma, Manik, Gurvinder Singh, and Rajinder Singh. "Design of GA and Ontology based NLP Frameworks for Online Opinion Mining." Recent Patents on Engineering 13.2 (2019): 159-165.
23. Cao, Ziqiang, et al. "Retrieve, rerank and rewrite: Soft template based neural summarization." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
24. Kumar, A. K. S. H. I., and A. D. I. T. I. Sharma. "Systematic literature review of fuzzy logic based text summarization." Iranian Journal of Fuzzy Systems 16.5 (2019): 45-59.
25. Kouris, Panagiotis, Georgios Alexandridis, and Andreas Stafylopatis. "Abstractive text summarization based on deep learning and semantic content generalization." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
26. Mosa, Mohamed Atef. "Data Text Mining Based on Swarm Intelligence Techniques: Review of Text Summarization Systems." Trends and Applications of Text Summarization Techniques. IGI Global, 2020. 88-124.

AUTHORS PROFILE

PL.Prabha, Research Scholar , Dept. of CSE, K.L.N. College of Information Technology,Pottapalayam ,Sivagangai Dist.

Dr.M.Parvathy,Professor,Dept. of CSE , Sethu Institute of Technology, Pulloor , Kariapatti.