Advances in Parallel Computing Technologies and Applications
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC210165

Effective Cataloging over Diverse Algorithms for Automatic Text Summarization and Its Survey

Pradheeba U [a,1], Sherin Glory J [b], Jausmin KJ [c] and Ramya U [d]

[a] *Department of CSE, R.M.K.College of Engineering and Technology, India*
[b] *Department of CSE, Rajalakshmi Engineering College, India*
[c] *Department of CSE, R.M.D. Engineering College, India*
[d] *Software Engineer, Techwirks, India*

Abstract. In these days, the measure of information and data accessible over the web is gigantic which prompted the making of automatic text summarization. Text Summarization is the route toward obtaining eminent information from a substantial text record. Automatic Text summarization is being a need of great importance and a fascinating theme with regards to NLP. Given the expansion in size and number of archives accessible on the web, an effective automatic text summarizer is significant. The fundamental test is to cause the computer to comprehend the report given with any expansion and create the outline as the primary rationale. The fundamental main impetus behind this work is to diminish the time and exertion spent by the client on perusing the whole document to understand what is the issue here. Subsequently the paper focuses on procedures accessible in delivering the significant rundown by utilizing different abstractive and extractive techniques

Keywords. Text Summarization, NLP, Text Classification, Deep Learning, Word Cloud.

1. Introduction

Over the years, there is a sensational development of the world's population. Consequently, it builds information on the web. The accessibility of the information is expanding step by step. Information can either be organized or unstructured, so the most ideal approach to explore is, look and decrease the outcome into shortening structure. So, there is an incredible need to abbreviate the text information by catching the salient features and this process is called Text summarization. The base pillar behind the Text summarization is Natural language processing(NLP) is a subfield of etymology, software engineering, and man-made thinking stressed over the associations among PCs and human language, explicitly how to program PCs to gauge and look at a great deal of basic language data [1].

[1] Pradheeba U, Department of CSE, R.M.K.College of Engineering and Technology, India.
E-mail: prathiba.ulaganathan@gmail.com.

NLP has had an uncommon development lately in the field of creation, one such innovation is automatic Text Summarization. It is comprehensively named Abstractive Text Summarization (ATS) and Extractive Text Summarization(ETS). Extractive text summarization is a clear method of shortening the original text content. The rundown is made by duplicating the pertinent sentence/words from the input text document though Abstractive Text summarization recreates significant context after understanding and assessment of the text utilizing advanced natural language strategies to produce a synopsis that passes on the most critical information from the original document [2].

This paper momentarily explains on various methodologies and strategies utilized in Text summarization and coordinated as follows. Section II portrays the elaborated literature survey; Section paper III gives the details analysis of the algorithms surveyed with a comparison table. Lastly, we concluded this paper with some future work that might be incorporated in our future work.

2. Literature Survey

Over the years there were numerous text summarization algorithms have been advanced. Each has its own advantages and disadvantages. In this literature survey, the most recognizable and broadly utilized summing up algorithms have been contemplated. Figure 1 below shows a summed-up block diagram for each one of those algorithms.

Anusha Pai [3] has proposed a framework for text summarization that is a combination of both statistical and linguistic examination of text reports. Linguistic summarizers use data about the language (grammar/semantics/use, etc.) to summarize a chronicle. Statistical summarizers work on word recurrence. The methodology has three fundamental parts: pre-processing, analysis, and selection. The framework has three segments: User who utilizes the framework, Summarizer which produces the outline, Database where there are isolated tables for putting away sentences, words, word recurrence, and sentence weight. Rundown age includes pre-processing, Word recurrence estimation, Plural resolution, Abbreviation resolution, Linguistic analysis, Sentence weight calculation, and normalization.

Outline created is superior to simple statistical summarizers that produce rundown dependent on word recurrence computation. The addition of plural resolution and abbreviation resolution adds more precision to the synopsis. Idea of standardization presented here causes sentences to get their weight simply dependent on the worth of its content words and not on the quantity of words it has.

Hongyan Jing [4] presents a novel sentence reduction framework for summarization purposes. The English Slot Grammar (ESG) parser is utilized to break down the syntactic construction of an information sentence and produce a sentence parse tree. In the reduction program, the parse tree is created which is annotated on with additional data. Grammar check is done dependent on basic, linguistic standards. The framework at that point distinguishes the words that are generally identified with the primary subject and the words that are connected to it and afterward processes a significance score.

The program utilizes corpus proof to figure how likely people eliminate a specific expression. In light of these various probabilities are processed: the likelihood that an expression is eliminated, the likelihood that an expression is decreased, and the likelihood that an expression is unaltered by any means. These corpus probabilities are computed beforehand using a training corpus. The system navigates the sentence parse tree and chooses which subtrees ought to be taken out, diminished, or unaltered. A subtree (i.e., aphrase) is taken out just in the event that it isn't syntactically compulsory, and it isn't the focal point of the local context. Dima Suleiman and Arafat Awajan [5] suggesed that text abstraction can be categorized into a few main classes dependent on genre, function, outline context, kind of summarizer, and the report style. This work actually gives an brief explanation about the methods, tuning metrics, datasets and the problem of DL dependencies over the abstractive Text summarization. DL endeavors to behave like what the human thinking can accomplish by taking out features at various levels of shrinking the text. Also Deep learning is been in application for a few language processing undertakings as it works with the observed facts of different level depictions of knowledge utilization in few processing multiple sections of non-linear units.

Pratibha Devihosur and Naseer [6] performed semantic unsupervised learning, so far and the effects are great as per them. The most important key take away in a sentence that belong to a information content is briefed by Lesk and wordnet, a language database for finding the connection among the words is efficiently utilized. Also this has been in Lesk's marking that it is been in connecting with every other word form to find the guarantee of every sentence. This is the most notable stage; texts are added with some prefixed weights and are managed in concaving requesting with their well-mean weights. All weighed out sentences are clearing to give meaning the total meaning as shown in the percentange of results.

Song et al. [7] carried out the Abstractive text summarization utilizing a deep learning system. In this system, the information is taken care of as a summary construction. This framework contains 3 stages information pre-processing, phrase processing, and text generation. The information pre-processing utilizes the core NLP to handle the passage since core NLP utilizes numerous linguistics tools. The processed text will be entered into the phrase process where phrases of the original summary are decreased the phrase decrease is finished utilizing Multiple Order Semantic Parsing (MOSP). The yield of the phrase cycle will be another phrase arrangement from which the text generation happens. The text summarization measure takes up the LSTM-CNN encoding and decoding method and, in this manner,summary is produced. Many key issues of text summarization are tackled in the LSTM-CNN model. The present ETS models are concerned about syntactic structure, while present ATS models are worried concerned semantics. ATSDL system outflanks the top tier models with respect to both semantics and syntactic development and achieves genuine results on manual linguistic quality assessment.

Nandhini and Balasundaram [8] have introduced a separate method called cohesive summary to monitor reading difficulties with the help of GA. The motive with this piece of work is to make the deletion over accurate mix of sentences that could grow thoughtful over the sentences and how they got bonding using GA. The information text is preprocessed using sentence segmentation, tokenization, stop words removal, and stemming is made to create feature extraction. The informative score, similarity of sentences, enhancement is made. Chromosome encoding is an important strategy in the genetic algorithm where chromosome refers to the order of sentences the chromosome size will be fixed and the summary will be compressed this in turn given to the fitness solution where fitness function, crossover mutation, and other works will occur. Finally, the summary will be generated. Genetic Algorithm based methodologies will be so helpful in learning how to do text abstraction for life science related texts. Also this helps in solving many legal statements to get summarized.

Deepa Nagalavi and Hanumanthappa [9] recommended multi-document summarization executed by a blend of qualitative and quantitative strategies on query-based. The input to the model is the articles and the query. It produces the summary in a few lines based on named entities. It works in three stages, first is named entity recognition model is utilized to distinguish and extricate the entities, then create the summed-up sentence utilizing entities as keywords and finally, joins the outline of various articles of a comparative topic. The query analyzer works by recognizing the key entities, which gives the data of area (event of an occasion) trailed by distinguishing the reliance or the relationship with previously recognized data. In the accompanying stage, by using the composed reliance strategy the reason for the event is perceived. Accordingly, the dependency relations and key entities are created by utilizing the Stanford parser [10] and regexner annotator [11]. Utilizing this data, a multi-document rundown is created in this methodology.

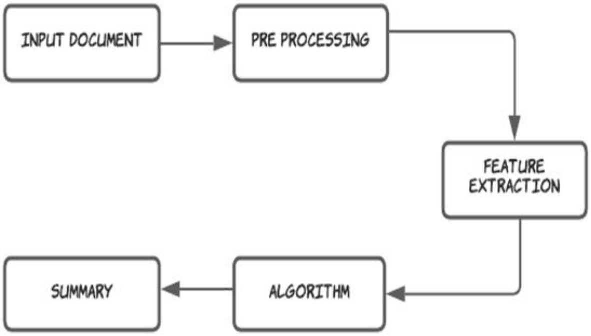

Figure 1. General Architecture Diagram for Text Summarization Algorithm

Prachi Shah and Nikita [12] proposed a model named Automatic Text Summarization Techniques for Indian and Foreign Languages. This is fundamentally centered around various languages in India. Each language has a different arrangement of documents for comprehension and summarization. It uses linguistic procedures to analyze and decipher the text and subsequently to find the novel thoughts and terms to best portray it by creating new more restricted text that passes on the main data from the original text document. The strategy it utilizes for all languages is text pre-processing, utilizing segmentation, tokenization, stop words removal and feature extraction is done. The sentences are ranked and the highest level are taken into consideration. It functions admirably for foreign dialects and automatic summarization for the Indian language is inadequate.

3. Analysis

So far in the past section, we have seen the different methodologies for Text Summarization algorithms, and now let us examine a few laurels about these algorithms and relative analysis about these algorithms. This segment additionally shows few applicable scenarios of these algorithms through Table 1 where these can put forth a valiant effort.

3.1. Single vs multi-document analysis

The automatic text summarization can be characterized dependent on the quantity of input documents. It is classified as single document text summarization and multi-document text summarization. In single document summarization, the synopsis is created from the single document though a multi-document summarizer contains numerous documents and delivers a single outline. Multi-document summarization is moderately more troublesome than single-document summarization as it includes composing numerous documents, extracting the significant context from each document, and delivering the cohesive summary.

3.2. Approaches

The accompanying methodologies have been recognized as the best methodologies under Extractive-based text summarization. The extract from the paper and the examination are given beneath.

3.2.1. Extractive based text summarization

Corpus-based Approach: Corpus linguistics is a quickly developing methodology that utilizes the statistical analysis of enormous assortments of composed or spoken information (corpora) to explore linguistic wonders. Corpus is a predefined assortment of words or wordnet [13] of different domains. The corpus-based methodology compares the sentence of the first content with the connected corpus and discovers the significance between them by utilizing the TF-IDF(term frequency-inverse document frequency). The most regularly utilized corpus in NLP is ”brown corpus” which comprises 500 English language text samples and 1 million words. This corpus is utilized for parts of speech tag(POS). For instance, the word “students” would have a set of words in the word net-like “school” “college” “books” etc.

Cohesion Based Approach: Corpus-based approaches neglect to represent relations between sentences in an archive. Text cohesion alludes to the connection between words that are utilized while creating a summary instead of just relying upon the corpus. It helps in creating significant and organized Graph-Based Approach: For the situation of graph-based methodology each archive or each sentence is portrayed as a node and the relation between them as edges. Edges are utilized to interface any two nodes sharing common data. Sentence scoring is finished by introducing weightage to the nodes of the graph. The likeness between those nodes is addressed as the edge score. For the most part utilized calculations are for the rundown. Cohesion is guaranteed utilizing lexical chains. These chains have related and subordinate words together in a solitary chain. These chains are then assessed and scored based on their sort and connection in the text.

Table 1. Analysis of various Text Summarization Algorithms

S No	Title	Author	Observation	Evaluation Metrics
1	Text Summarizer Using Abstractive and Extractive Method	Ms. Anusha pai May 2014	The proposed approach culminates both linguistic and statistical analysis which provided an efficient result	The document containing 10 sentences (sentence 1 to sentence 10) and summary size is 40% in this approach
2	Sentence Reduction for Automatic Text Summarization	Hongya Jing 2016	The system uses a parse tree to find the dependency between words and take reduction decisions.	The program achieved an average success rate of 81.3% using anfivefold approach
3	Deep Learning Based Abstractive TextSummarization: Approaches, Datasets, Evaluation.	Dima Suleiman and Arafat Awajan Aug 2020	The abstractive summarization using Deep learning techniques and evaluated against various measures.	Quantitative approach (ROGUE) and Qualitative approach(Human evaluation) is used to evaluate the summary which ensures readability, relevancy and fluency.
4	Automatic Text Summarization using NLP	Pratibha Devihosur and Naseer R Aug 2017	LESK algorithms did its excellence in this approach and word net corpus was used.	ROGUE method is used to validate the summary generated by the system and the human generated summary
5	Abstractive Text Summarization using Deep Learning	Shengli Song Haitao Huang Feb 2018	Text summarization using ETS based on deep learning techniques such as LSTM,CNN,Seq 2 Seq model.	In this approach ROGUE method is used to evaluate the summary and it has become successful.
6	Use of Genetic Algorithm for Cohesive Summary Extraction to Assist Reading Difficulties	K. Nandhini and S. R. Balasundaram June 2013	The main purpose is to take out the possible combination of sentences that could not only increase readability but improves the bonding over the related.	Intrinsic method is used to evaluate the precision, recall and F-measure while extrinsic method is used to evaluate the readability and usability of the summary generated.
7	The NLP Techniques for Automatic Multi-article News Summarization Based on Abstract Meaning Representation	Deepa Nagalavi and M. Hanumanthappa Nov 2013	Combined approach of qualitative and quantitative on query based using regexnerannotater and stanford parser.	The proposed approach is evaluated by employing both qualitative and quantitative evaluation metrics.
8	Automatic Text Summarization Techniques for Indian and Foriegn Languages	Prachi Shah and Nikita P. Desai 2016	It emphasizes all about the survey and performance analysis of different language.Automatic summarizer is more successful for foreign language than compared to indian language.	Initially manual testing was done. Later LibSVM was used for evaluation and accuracy of 75% was achieved.

These scores are subsequently used to produce the genuine summary. For the most part utilized calculations for the graph based-approach are the Text-Rank and Google's

Page-Rank calculation [14]. Text-Rank calculation ascertains the sentence similitude based on the TF-IDF mode and the sentences are ranked while Page-Rank is utilized to compute the importance between the reports under comparable substance. The graph-based methodology has been the productive methodology for extractive-based text summarization as the summary generated is more cohesive and based on both frequency and similarity.

3.2.2. Abstractive based text summarization

Seq2Seq Model: Seq2Seq is utilized to take care of issues on sequential data. The abstractive summary can be produced utilizing a many-to-many seq2seq model where the input is a long sequence of words and yield is a summary. This model has two parts specifically i) Encoder ii)Decoder each of which is an RNN. The encoder peruses a single word per timestamp and measures the word. It at that point catches the relevant data present in the input sequence. The decoder examines the entire target progression word-by-word and predicts a similar arrangement balance in one timestamp. The decoder is an idea to anticipate the next word in the sequence, given the previous work. In any case, this model has the restriction over the more drawn-out sequence of words as it is hard for the encoder to remember long sequences into a fixed-length vector [15]. Attention Mechanism: Considering the impediments in the Seq2Seq model, attention mechanisms have come into play. It plans to foresee a word by taking a gander at a couple of explicit pieces of the sequence instead of the whole sequence. The context vector can be determined by utilizing global attention as well as local attention [16]. Global attention considers every one of the secret states though Local attention thinks about just the chosen covered up states. Consequently, it is utilized particularly when the sequence is enormous. In light of the examination, a framework can be proposed as demonstrated in Figure 2 by bunching the articles, and keywords are extricated from each bunch utilizing topic modeling at that point by getting an information keyword from the client the pertinent group and applicable articles are gathered and a multi-record synopsis is produced. Consequently, for any keyword given, an outline is produced.

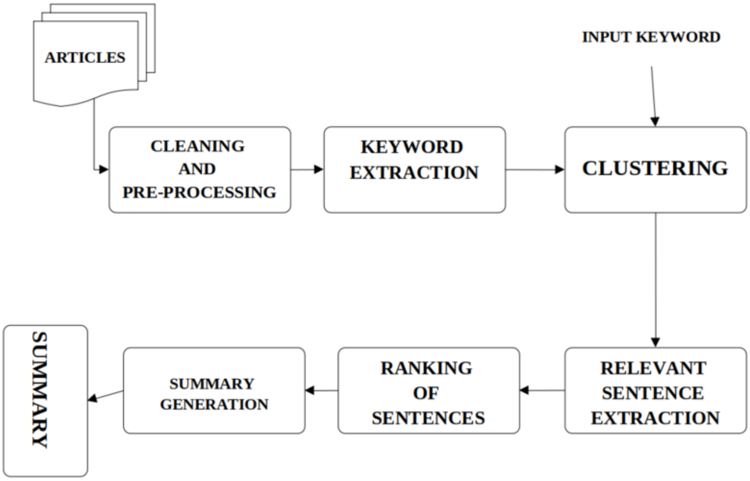

Figure 2. An Improvised Architecture Diagram for Text Summarization

4. Conclusion

In this paper various techniques used for text summarization were discussed. It can be concluded that an extractive summary can be generated on any domain irrespective of the corpus available but it doesn't provide an anthropocentric summary, whereas abstractive summary is more of anthropocentric but limits to the corpus available. Hence the combination of both extractive and abstractive techniques can be incorporated in producing a more meaningful summary. This will be incorporated in our future work and according to the model architecture depicted above a model will be implemented with different data size and various evaluation metrics.

References

[1] Wikipedia, The Free Encyclopedia. Natural language processing [Internet]. [updated 2021 Mar 25; cited 2021 Mar 30]. Available from: https://en.wikipedia.org/wiki/Natural_language_processing.

[2] Wikipedia, The Free Encyclopedia. Automatic summarization [Internet]. [updated 2021 Mar 9; cited 2021 Mar 30]. Available from: https://en.wikipedia.org/wiki/Automatic_summarization.

[3] Pai A. Text Summarizer Using Abstractive and Extractive Method. International Journal of Engineering Research & Technology. 2014;3(5):2278-0181.

[4] Jing H. Sentence reduction for automatic text summarization. InSixth Applied Natural Language Processing Conference 2000 Apr (pp. 310-315).

[5] Suleiman D, Awajan A. Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures, and challenges. Mathematical Problems in Engineering. 2020 Aug 24;2020:202-9.

[6] Devihosur P, Naseer R. Automatic text summarization using natural language processing. International Research Journal of Engineering and Technology (IRJET). 2017;4(08):667-73.

[7] Song S, Huang H, Ruan T. Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications. 2019 Jan;78(1):857-75.

[8] Nandhini K, Balasundaram SR. Use of genetic algorithm for cohesive summary extraction to assist reading difficulties. Applied Computational Intelligence and Soft Computing. 2013 Jan 1;2013:412-20.

[9] Nagalavi D, Hanumanthappa M. The NLP Techniques for Automatic Multi-article News Summarization Based on Abstract Meaning Representation. InEmerging Trends in Expert Applications and Security 2019 (pp. 253-260). Springer, Singapore.

[10] De Marneffe MC, Manning CD. Stanford typed dependencies manual. Technical report, Stanford University; 2008 Sep:880-8.

[11] Sarkar D. Semantic analysis. InText Analytics with Python 2019 (pp. 519-566). Apress, Berkeley, CA.

[12] Shah P, Desai NP. A survey of automatic text summarization techniques for Indian and foreign languages. In2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) 2016 Mar 3 (pp. 4598-4601). IEEE.

[13] Kulshreshtha RK, Srivastava A, Bhardwaj M. A Survey Paper on Text Summarization Methods. International Research journal of Engineering and Technology (IRJET). 2018 Nov;5(11):97-104.

[14] Moawad IF, Aref M. Semantic graph reduction approach for abstractive Text Summarization. In2012 Seventh International Conference on Computer Engineering & Systems (ICCES) 2012 Nov 27 (pp. 132-138). IEEE.

[15] Pai A. Comprehensive guide to text summarisation using deep learning in Python. Blog, Analytics Vidhya, June. 2019;10:456-462.

[16] A Comprehensive Guide to Attention Mechanism in Deep Learning for Everyone. American Express. 2019 Nov:557-63.