e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:06/June-2022 Impact Factor- 6.752 www.irjmets.com

A REVIEW ON ABSTRACTIVE TEXT SUMMARIZATION

Bhupendra Singh Patel[*1]

[*1]Student, Department Of Computer Science, MITS, Gwalior, Madhya Pradesh, India.

ABSTRACT

Text summarisation is the method of converting a long text into a shorter text which preserves the meaning of original text, it includes all the main points contained in the text into a continuous readable text. In the world of internet information is increasing like crazy it may be in the form of images, videos, audios, text blogs. It is very hard and time-consuming process to find out the relevant information of our need. Here comes the process of automatic text summarisation (ATS) which can be very useful in this case. By summarisation of information, we can easily analyse it, share it, and use it. Summarisation of text can be of two types Extractive and Abstractive. In this paper we talk about the Abstractive way of summarising things, it is way in which we reduce the text in a way that it contains all the important text is about preserving its meaning, it did not contain the exact sentences used in the text but modified text. This way summarisation things can be very useful in the field of creating short news, sending compressed websites, tech articles review and generating automatic abstract of research papers. This paper focuses on the semantic analysis and idea behind summarisation technique. It sums up with the future usage of this technique.

Keywords: Abstractive Summarisation, Semantics, Extractive Summarisation, Compressed, ATS.

I. INTRODUCTION

Data is like floating around us, it's everywhere, everything is running on the data these days. Data is generating at a rate which is causing difficulties to store it. Data is stored in the form of a binary numbers in semiconductor or magnetic devices, it is costly to store and manage this data. Companies now use data as the fuel, their decisions depend on the data analysis of previously used data. This Data contains information that is useful and also trash information that just taking space to store. It is very difficult to find out which information is useful and which is not. The source of this information is internet stored in the form of images, video, audio, graph, and tables. Data content is stored in very unorganized manner making it difficult to filter and utilize the prewritten content. To make this available information useful we have to classify it after shortening it make it usable. This content can be summarized by two ways first one is Extractive technique another one is Abstractive Technique. Extractive technique basically selects the important sentences from the text make it short. Extractive very old technique used for document summarization. It is similar to that student used to mark the important text from books for later references.

In the corpus based checking the whole text is scanned for the important information in it, then this information is used to check the similarity among the words, text. Semantic similarity can be used in the Natural Language Processing to perform task related to information optimization, redundancy removal from the text document [1]. Abstractive summarisation is the way in which text is shortens in a way that main idea of the text is preserved but the same sentences used in text are not present. Abstractive summarisation is attracting researchers because of having potential to generate novel term using the Natural Language Processing. This is pretty much useful in creating new articles from existing ones. Its best use occurs when unorganised blog when processed can give a new perfect blog. Researchers uses this technique to quickly extract useful information from the research papers. This technique works on the bases of semantics classification, semantics frequency helps the tokens to rank in priority queue so that summarisation can be done according to the word limit. Abstractive technique uses the semantic analysis to summarise the text. Semantic similarity measurement is term associated with the similarity between the actual meanings of the word itself, Natural Language Processing achieves this by the help of words represented in term of vectors. Vectors are the mathematical representation of word; this allows to check for the semantic similarity. Similarity between semantics can be checked using two methods first one is knowledge based checked and another is corpus based. In the knowledge-based checking it manually checks the meaning in sematic database.

II. WORK RELATED TO ABSTRACT SUMMARISATION

In the paper [2] discussed that the measurement of similarity between words on the basis of distance between the two words. If two words appear together it implies zero distance. Distance value can be differ based on the closeness of the words. Using frequency-based approach, semantic similarity can be calculated. Words appearing higher number of times should be assigned with higher score. Another method is to find the semantic similarity is by checking neighboring words in semantic database. calculation of similarity is done on the repetitive occurrence of the neigh boring words which belongs to have same meanings. Normalized distance is defined for semantic similarity, which is determined by checking the word with the group of keywords. Two words are said to be close if they have the same meaning. Words which are present side by side have distance zero and constant is defined for the word which are not together [1].

Abstractive text summarisation mainly classified into following category:

(1) Predefined Structure Based

This method uses predefined frameworks to find out significant words from the text that make up summary. These frameworks are Tree, Graph, Template and Ontologies [3].

(2) Semantic Analysis Based

This method uses word converted into semantics, semantics than analysed using natural language generation (NLG), semantic picture of the text document is created. NLG algorithm uses semantic-graph, data-items and predicate opinion to create semantic picture of text, then this picture is use NLG to generate abstract summary [3].

(3) Deep Learning Based Method

lin and ng (2019) demonstrated another way of summarising things a technique based on neural networks [4] neural used to generate the summary.

1.1 Graph Based Approach

A technique known as "Opinosisi", it is developed on graph also called as text rank algorithm in which every word is node of the graph, where the direct edge between the words represents the similarity between word. If the similarity between phrases found more than some predefined value a direct edge is created between them.

1) Construction of the graph: word-based graph is generated to perfectly describe the source

2) Summary Creation: The process in which the output abstractive summary created. Many sub-methods of graph are being found out and described as follows:

1. A vote is assigned to every node vote increases when a node connects with it, these votes sorted out in descending order, higher the vote number more important is the word. Unused paths score also includes in process of sorting to make to summary more accurate.

2. By applying likeness measure (e.g., Jaccard) redundant or very similar path can be removed.

3. After step two topmost part of the remaining paths consists of summary, length of summary is dependent on the number of graph paths picked.

1.2 Graph Based Approach

This method founds comparable sentences that exchanges information between them, then collecting these sentences to generate summary [3]. Equal sentences are represented by a tree structure. Through parsing the dependency tree is constructed. To describe the text document in the form of tree, the tree-based method is often used. In procedure to produced summary, some tasks are progress like trees pruning and linearization (i.e., translating trees to strings), etc. [3]. Multi-document abstract summarisation was demonstrated by Kurisinkel, Zhang, and Varma [5]. The steps used in this approach are as follows:

1) input text of the corpus is parsed to get all the sets of Syntactic dependencies.

2) Retrieving all the unfinished dependency trees of different sizes from trees formed in the first stage

3) Clustering the chosen unfinished dependency trees from all sets of trees to generate summary in the range.

4) using the clustered tree to generate the sentences it shows the important of clustered tree to generate the summary.

1.3 Rule-Based Approach

In this method some rules and classes of methods are written to found the necessary ideas about the input document to generate the summary. Following steps are used in this method are [3]:

1) Input document is classified on the basis of prewritten rule that found out the relationship and meaning behind the words.

2) A query is formulated with respect to the input given.

3) Queries are resolved by finding the relationships and main ideas of the document.

4) these responses are passed into outlines rules to generate the abstractive summary of document.

Genest and Lapalme (2012) [6], proposed a style which works upon the abstractive structures. Each abstractive structure is planned to solve a smaller text part or ideas which can be easily solved. These proposed rule-based approach can be used in smaller application in future. IE (Information Extraction) and simple patterns creation are used to create pattern for every word structure. All these guidelines generated manually. An abstractive system looking to response for single or multiple features which could be linked with the equal feature. The Information Extraction guidelines might be customized according to applicants for every feature they need which will involve in summary creation task.

1.4 Semantic-Based Approach

A approach in which predicate argument data objects, or semantic graphs makes a semantic image of text which then passed this information to Natural Language Generation (NLG) where verb and nouns combined to make the abstract summary of the text. According to [6] represent input text in the form of a predicate argument structure, these structures processed to form the clusters generated by checking semantically equal predicate from the text. Now on the basis of weighted and optimised grade predicate structures are generated, final summary can be created using these predicates. Drawback of this method is that the summary is constantly checked by human which can be automated [7].

1.5 Deep Learning Based Approach

Using sequence to sequence (seq2seq) method an near perfect abstractive summary can be generated shown in the [8]. Seq2seq is a established model over several Natural Language Processing in the fields of object recognition, tracking, Machine learning and in chat bots. A recurrent neural network model constructed on the technique of encoding-decoding works perfect on the short text but unable to fetch same results when applied for the longer texts. This drawback arises due to the limited vocabulary of deep learning module. Deep learning model has some of disadvantages like repetitive words or sentences arrives frequently. Have very less capability of handling infrequent words and out of the vocabulary words. This model has very less disadvantages in order to overcome these, some methods can be applicable first, we can shorten the original text by pre-processing, dividing it into the shorten texts and taking the summary of these separately and combing these short summaries to make for original one. Second method by using word vectorisation and implementing it in the pre-trained model and using it in original model. Third method is using the TensorFlow model using bidirectional LSTM to encode the transcript and then decoding using a unidirectional LSTM method. With the summarisation, analyse the noise produced try to optimize it.

III. CHALLENGES

Redundancy- redundancy always brings bad effects on the summarised documents. It doesn't sound good to read document filled with duplicity. If redundancy can be removed it can fit more content in that matter only. Similarity measurement is the most important task of the summarisation by which it can be reduced. If we precisely measure the duplicity, it can be found and reduced.

Irrelevancy- main motive behind the summarisation is that we want to remove irrelevant information, making the content crisp and clear. Summarisation creates problem of hallucination means sometimes the summarised text contains the word which is not present in the text, and are irrelevant to the original text. This problem occurs because it is difficult to incorporate all the method in a given ratio such that it always produces a good relevant text. It is crucial to know all the features we want in the summary, it should be written according to the need of database, by this way we can reduce irrelevancy.

Loss off Coverage- the main aspect of the summary is that want to cover all the important points of the text. A good summary depicts every useful information given in the summary. Current summarisation technique does not focus on covering the whole things, so often fails to generate a good summary. This problem arises when to summarize multi-documents the number of topics is more than the single document. There are some techniques of literature summarisation, which focuses on covering things but this also causes redundant information. If we go for removing redundancy it covers less and vice versa.

Non readability and less cohesiveness- summary should be readable and continues. By readability and cohesiveness means that content should be conceptually related, it doesn't feel like some of the part is missing in between. There should be a index that can show the cohesiveness of the content.

IV. CONCLUSION

There have been many applications of natural language processing and automatic text summarisation is one of the popular techniques. There are basically 2 ways to get text summarization one is extractive summarisation and other is abstractive summarisation. Research in the field of automatic text summarisation is happening from very early days. Now days researchers are more focusing on abstractive summarisation instead of extractive summarisation. Automatic text summarisation technique generates relevant, content that depicts same meaning with fewer redundancies. Our goal in this field is to reach summarisation nearer to the human intelligence it's tough work but things are improving day by day. Therefore, this study gives insight of the techniques used for Abstractive summarisation with their advantage and disadvantage with respect to each way. Major techniques used are discussed. This research gives insight on the abstract summarisation and its ways.

V. REFERENCES

[1] Supreetha. D1, Rajeshwari. S. B2, Jagadish. S Kallimani3 Student1, Assistant Professor2, Associate Professor3 "Abstractive Text Summarization Techniques" Ramaiah Institute of Technology, Bengaluru, India published at IJESC journal, 26884-26888, 2020

[2] G. Silva, R. Ferreira, S. J. Simske, L. Rafael Lins, M. Rissand H. O. Cabral, "Automatic text document summarization based on machine learning," pp. 191–194, 2015.

[3] Gupta, S., & Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. Expert Systems with Applications, 121, 49–65. https://doi.org/ 10.1016/j.eswa.2018.12.011.

[4] Lin, H. & Ng, V. (2019). Abstractive summarization: A survey of the state of the art. Paper presented at the The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI- 19).

[5] J Kurisinkel, L., Zhang, Y. & Varma, V. (2017). Abstractive multi-document summarization by partial tree extraction, recombination and linearization. Paper presented at the Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Taipei, Taiwan.

[6] Genest, P.-E. & Lapalme, G. (2012). Fully abstractive approach to guided summarization. Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers – Volume 2, Jeju Island, Korea.

[7] Atif Khan and Naomaisalim, "A review on abstractive summarization methods", 2014.

[8] Khan, A., Salim, N., & Jaya Kumar, Y. (2015). A framework for multi-document abstractive summarization based on semantic role labelling. Applied Soft Computing, 30, 737–747. https://doi.org/10.1016/j.asoc.2015.01.070.