IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10402

Where are the Automatic Text Summaries Located in the 2021? A review

Jhonathan Quillo-Espino[1], Rosa-María Romero-González[2]

Assistant professor, Faculty of Computer Sciences, Autonomus University of Queretaro, Santiago de Queretaro, Queretaro 76230, Mexico[1]

Professor, Faculty of Computer Sciences, Autonomus University of Queretaro, Santiago de Queretaro, Queretaro 76230, Mexico[2]

Abstract: The creation of automatic text summaries is the result of the execution and application of text mining. Its complexity consists in finding the most relevant information, different methods or perceptions have been developed through time for this matter. Research shows a detailed analysis of the execution, presenting the most representative characteristics of the automatic extractive summaries generation, being this one, one of the methods developed in the recent years.

Keywords: Text Mining, Extractive Text Summarization, Abstractive Text Summarization, Information Retrieval.

I. INTRODUCTION

The struggle researchers have encountered to find new development technologies that allow the improvement of the understanding processes and changes of the natural language processing (NLP) [1] declare that is the assisted technique done by the computer intended to analyse and comprehend human language. Computer language moves forward at a fast rate due to the interest in improving processes, in addition; it has a rapid development because of the limitless rise of internet texts, documents, web pages, bank records, research articles, etc. Nowadays computer science moves forward at an accelerated pace. To set a clear example, there is the artificial intelligence with different algorithms, neural networks, tree data structure theories, etc. In this research there is an exploration to the process of automatic text summaries. The ability to summarize was considered a natural task done by humans through reasoning and understanding of the information obtained from a reading, however; today it is an easy task with the help of a computer.

[2] Propose that a Text Summarization (TS), is a reduction of data for the user consumption. It is essential to understand the main idea of an Automatic Text Summarization (ATS), [3] determine that, essentially, it tries to condense or reduce a document or a group of documents while preserving the most relevant information and content without losing its meaning. The summary must contain the foremost information from the original text. The considerable and available amount of information that exists generates the need to synthetize it maintaining the importance and integrity of the meaning, moreover, the need to recover and subtract new knowledge. One of the most important arguments to support the existence of ATS are: the preservation of the most relevant information and the reduction of reading time. Another important factor about ATS is the advantage of providing a wide vision to the reader while giving a clear idea from a large text and with minimum effort. A limitation of ATS is that it needs digital texts, as well as certain characteristics such as text configuration in order for the computer to be able to read it and analyse it. If large data quantities of TS are done manually, it becomes a difficult and tedious task, in addition; it affects the results obtained from the reading texts.

The development and advances in managing information techniques have evolved granting technologies related with decision making with such information like Deep Learning (DL), pretending to duplicate the behaviour of the human brain towards the analysis of information to select words and generate new sentences, making possible the creation of ATS in a fast and efficient way. Besides, they can be applied to a single document or to many documents depending on its original resource.

II. GENERAL PROCESS OF ATS

The ATS are part of the Information Retrieval (IR) [4], [5], point out that ATS is in charge of recognizing and obtaining resources and relevant information for a specific need within large collections of information, which origin derives from Text Mining (TM), [6], define it as the one in charge of finding patterns and to discover hidden knowledge in the information. Fig 1 shows the general TM process.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10402

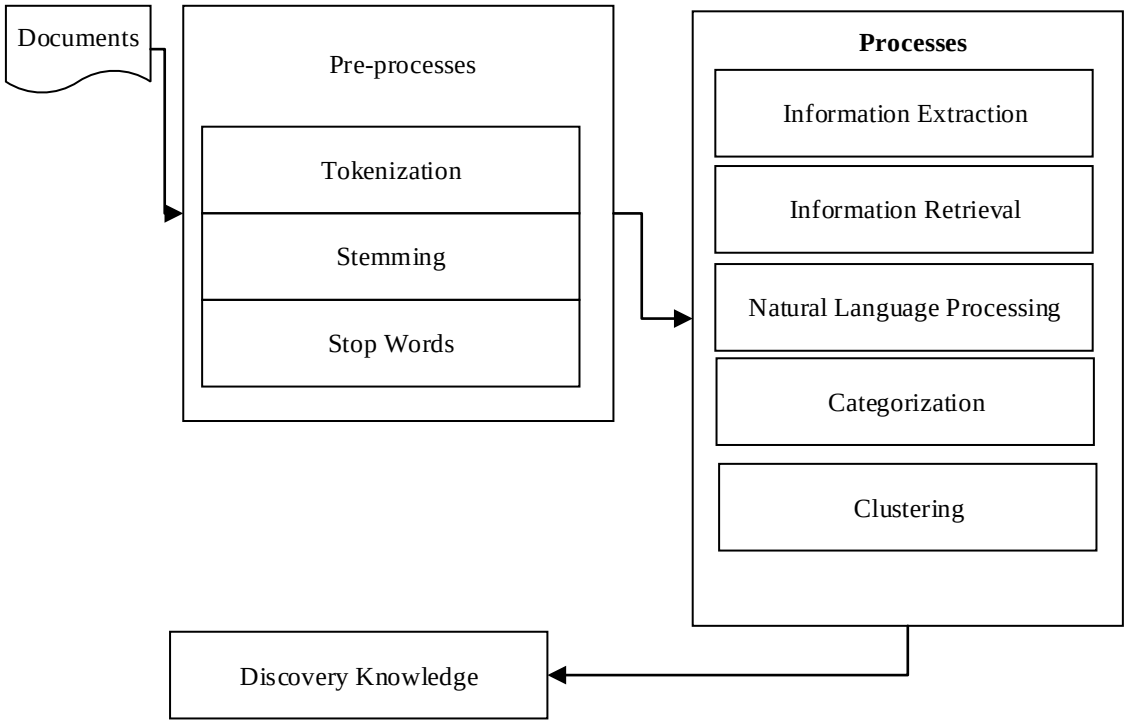

Fig. 1 General TM process [6]

The general ATS process derives from the execution of IR, which at the same time is part of the TM processes. The general process consists in processing the information which objective is to organize and structure the information in order to be used through the application of a method. Fig 2 shows the general RTA process.

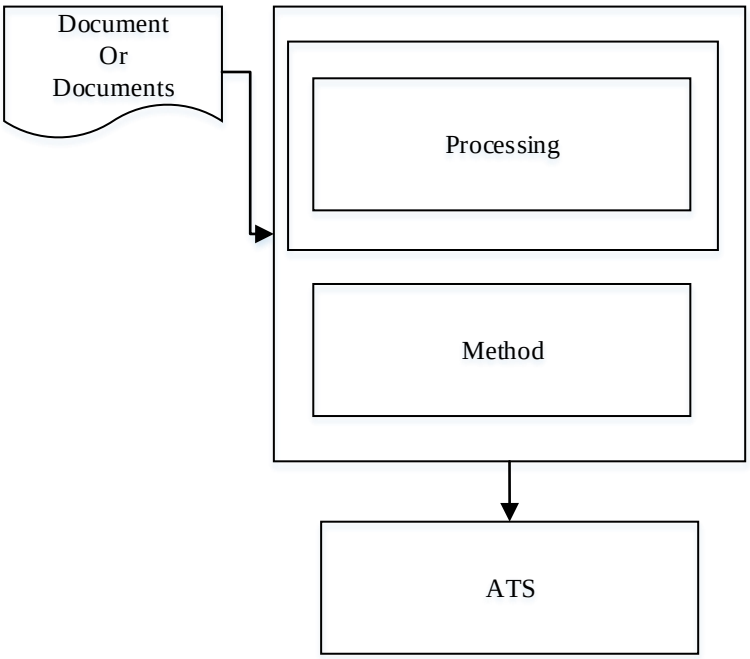

Fig. 2 General ATS process

ATS is a complex process for humans and machines and it was proposed by [7], a great variety of methods have been implemented using characteristics related with frequency, including phrases, words and keywords, jests, etc. For more than 6 decades, the amount of proposals has increased by researchers in charge to make text evaluation due to the massive growth of available information.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10402

III. ATS TECHNIQUES

There are two techniques mainly used for ATS. The abstractive text summarization approach (AATS) [8], propose that it summarizes text through the development of new sentences formed from the most relevant content of the text. The main idea consists in the automated and deep linguistic analysis of the text with the purpose to find concepts and ideas that allow the creation of short texts making a summary, it is a more complex process due to the number of tasks that need to be done. On the other hand, the extractive automatic text summarization approach (EATS) [9], state that it is the simple and sturdy way to generate ATS by selecting more representative sentences. The document is divided into different sentences, each sentence is analysed, counted, furthermore ranked by punctuation, with the sentences containing the highest punctuation being presented to the user. [10], mention that one of the advantages of EATS is the comparative simplicity of development as it does not require the creation of databases or advanced knowledge in linguistics for text detailing. Both approaches are part of text mining (TM) and the information retrieval (IR) and information extraction (IE). Fig 3 shows the EATS process using the TF-IDF method.

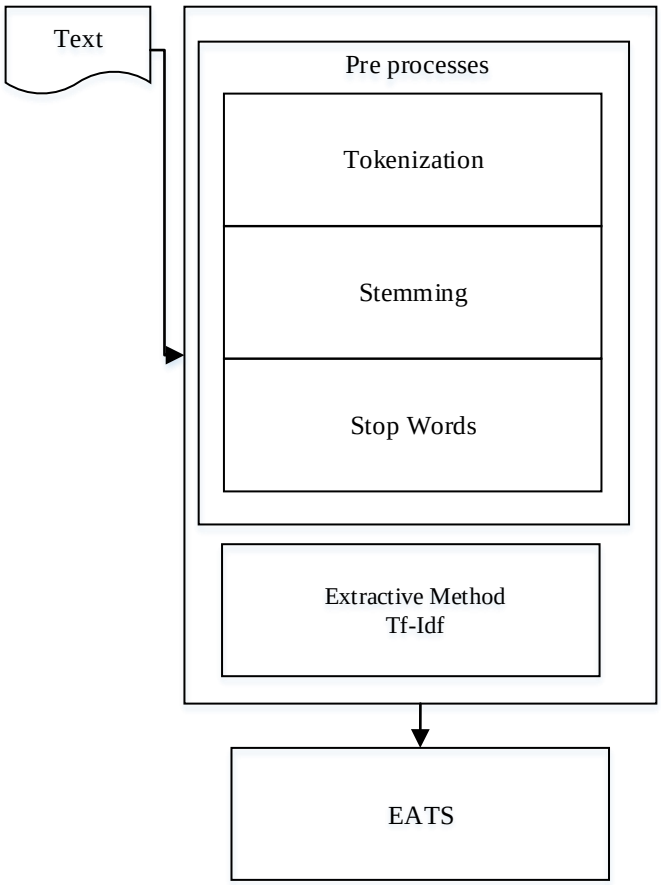

Fig. 3 General ATS process

The first pre/process is Tokenization [11], [12], concludes that is the process to identify units into the unique elements called tokens and to remove empty spaces. It collaborates to be able to perform the elements counting as combined cells. Table I shows an example of the tokenization process in Spanish.

TABLE I TOKENIZATION EXAMPLE

Data Entry	Output	Output amount of tokens
Los niños juegan en el patio	“Los”, “niños”, “juegan”, “en”, “el”, “patio”	6

By pulling apart the content in the elements, it permits to perform the second element in a faster and easier way. The second element called Stemming [13], [14], conclude that is the pre-process used to transform different words into their root-based form. It allows to find similar tokens for subsequence computation. Table II shows the Spanish stemming example.

TABLE II TOKENIZATION EXAMPLE

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10402

Data entry/Sentence	Output/Stemming
Los niños juegan en el patio.	“los”, “niñ”, “jueg”, “en”, “el”,“pati”

The third element called Stop Word [15], refer that they are tokens that appear frequently and commonly provide little analytical meaning and value in TM. Table III shows an example of a stop word list in Spanish.

TABLE III STOP WORD LIST IN SPANISH EXAMPLE

a, al, algo, algunas, algunos, ante, antes, como, con, contra, cual, cuando, de, del, desde, donde, durante, e, el, ella, ellas, ellos, en, entre, era, erais, eran, eras, eres, es, esa, esas, ese, eso, esos, esta, estaba, estabais, estaban, estabas, estad, estada, estadas, estado, estados, estamos, estando, estar, estaremos, estará, estarán, estarás, estaré, estaréis, estaría, estaríais, estaríamos, estarían, estarías, estas, este, estemos, esto, estos, estoy, estuve, estuviera, estuvierais, estuvieran, estuvieras.

Basically, it consists in eliminating tokens that are similar to the previously predetermined list. Table IV shows an application example of stop word list.

TABLE IV APPLICATION OF STOP WORD LIST: EXAMPLE

Date entry/ Sentence	Output/ Stop Words
los niñ jueg en el pati	“niñ”, “jueg”, “pati”

In Table 4, it can be seen the elimination of words leaving only as output a total of 3 words out of 6. Finally, the application of the selected method is executed.

IV. SUMMARY OF EXTRACTIVE AUTOMATIC TEXT (EATS)

Different statistical methods for the development or EATS which are mentioned below: Table V shows different methods for a EATS based on key phrases and frequency balancing.

TABLE V METHODS TO OBTAIN EATS

Method	Description
Localization Method	[16], indicate that it focuses on the origin of the position of the important information in a sentence. It can be at the beginning or the end of the sentence.
Cue Method	[17], determine that it is based on the hypothesis of which the relevance of the sentence is measured by the presence or absence of specific keywords coming from the word dictionary.
Heading Method	[18], state the sentence weight is calculated adding the number of words in the heading and compared with the rest of the titles.
Frequency Method	[19], conclude that provides the weight based on the frequency of each sentence element. The score increases or decreases depending on the number of the sentence. Commonly, the frequency term (FT) and the inverse document frequency (IDF) are used. The (FT-IDF) is a numeric statistic that projects the importance of a word for each document in the collection.
Sentence Length	Base on the length of the sentence
Thematic words	[20], refer that are the ones in which a word appears more often in the text, they are considered as thematic words if they are in the first 5 most repeated words.
Keywords or Noun Phrases	[21], states that they are plain phrases that include a noun before or after a word that describes the noun.
Linguistic Method	It is in charge of the semantic analysis and the language properties, as well as the inference between the meaning and the concepts with the text.
Graph Method	[22], deduces that represent the sentence as a group of words and they use a similar content measure that can detect redundant, semantically or equivalent sentences.
Cluster Grouping Method	[23], propose that it is based on grouping similar textual units such as paragraphs, sentences in order to identify common information topics.
Machine Learning (ML)	[24], define that is the one that uses certain characteristics extracted from original text, such as the frequency of some text elements and linguistic. The algorithms Naïve Bayes, tree decisions, etc., use learned patterns to classify the information and create EATS

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10402

Latent Semantic Analysis	[25], ratify that it is a statistical, algebraic method in charge of finding and extracting hidden semantic structures. It uses common words found in different sentences, a high number of common words among the sentences indicate that the sentences are semantically related.

V. ABSTRACTIVE AUTOMATIC TEXT SUMMARIZATION (AATS)

Below, some methods for AATS are mentioned. Table VI shows types of methods for AATS.

TABLE VI METHODS TO OBTAIN AATS

Method	**Description**
Ontology	[26], conclude that are the ones used in particular domains and are used to retrieve information, specifying the concepts from that domain through the ontology. One of its advantages is to share the common understanding of the structure among the members of the structure, moreover it includes aspects of time, periods of time and relative measurements.
Herarchical deep neural network	[27], propose a new neuronal structure similar to the human one to improve the performance of AATS, trying to replicate the human common sense by attempting to capture relevant document information at different levels, using text categorization to identify remarkable information while the syntax annotation tries to reduce the creation of grammar errors.
Templete based Method	[28], determine that is the technique applied to represent the entire document in a template, where the linguistic patterns are used to identify text fragments saved in the template. Such fragments are gauges of the summarized content, the advantage is a very coherent summary as a result.
Rule based Method	[29], define that a guide is used for the representation of the entire document, the purpose is to identify text fragments that can be mapped to generate summaries.

VI. CONCLUSIONS

ATS should be consider as tools derived from TM, its main function aids to read a text in a more efficient manner, reduce the reading time and at the same time to discover the most relevant information in a text. It helps to reduce the amount of information contained in a document, it can be applied to a document or to a set of documents, the complexity depends on the type of method. Semantics is essential for NLP and to be able to translate the text into computer language.

The EATS with FT-IDF method aim to divide the text into independent sentences to be processed and obtain a score of the sentences according to mathematical operations derived from statistics. Each sentence obtains a score level, then, they are arranged and compared with the required summary percentage level and the closest ones are obtained according to the threshold percentage. On the contrary, the AATS which main function is to generate a summary through new sentences using ontologies, neural networks, etc., to attempt to interpret the information and pretend to create new shorter sentences with the same meaning.

The ATS are excellent tools while detonating the usefulness of the TM nowadays, granting the gathering of information. Day by day the necessity to create ATS increases due to the high increase of information from enterprises, government and private scopes.

This research aims to aid the understanding of new research concepts in the area by offering a summary of the main and current techniques used in automatic text summarization methods. An example of the pre-process used in the performance of automatic extractive summaries is mentioned.

REFERENCES

[1]. Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, Hefu. (2020). REVIEW Natural language processing (NLP) in management research: A literaturereview. Journal of Management Analytics. https://doi.org/10.1080/23270012.2020.1756939

[2]. Mohd, M., Jan, R., & Shan, M. (2019). Text document summarization using word embedding. Expert Systems with Applications. Vol (143). https://doi.org/10.1016/j.eswa.2019.112958.

[3]. Mudasir, M., Rafiya, J., & Muzaffar, S. (2010). Text document summarization using Word embedding. Expert systems with aplications. ISSN 0957-4174. .https://duc.nist.gov/duc2007/

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10402

[4]. Salloum, S. A., Mostafa Al-Emran., Monem. A. A., & Shaalan, K. (2018). Using Text Mining Techniques for Extracting Information from Research Articles. In book: Intelligent Natural Language Processing: Trends and Applications, pp. (373-397). https://doi.org/10.1007/978-3-319-67056-0_18
[5]. Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., Wu, Ch., Croft, W. B., & Chen, X. (2019). A Deep Look into neural ranking models for information retrieval. Information Processing & Management. 57(6). https://doi.org/10.1016/j.ipm.2019.102067.
[6]. Quillo-Espino, J. Romero-González, R. M., & Lara-Guevara, A. (2018). Advantages of using a spell checker in text mining pre-processes. Journal of computer and communications. 6(11). DOI: 10.4236/jcc.2018.611004
[7]. Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159–165.
[8]. Huang, D., Cui, L., Yang, S., Bao, G., Xie, K. W., Zhang, Y. (2020). What we achived on text summarization. Conference on Empirical Methods in Natural Language Processing, pp. (446-469), 16(20). Association for Computational Linguistics.
[9]. Bhargava, R., & Sharma, Y. (2019). Deep extractive text summarization. International Conference on computational intelligence and data science. 167(2020), pp.138-146. https://doi.org/10.1016/j.procs.2020.03.191.
[10]. Batura, T., Bakiyeva, A., & Charintseva, M. (2020). A method for automatic text summarization based on rhetorical analysis and topic modeling. International journal of computing. On-line ISSN 2312-5381.
[11]. Hickman, L., Thapa, S., Tay, L., & Cao, M., & Srinivasan, P. (2020). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations In-press at Organizational Research Methods. DOI: 10.1177/1094428120971683
[12]. Tavana, M., Shaabani, A. Santos-Arteaga, F.J., & Vanan, I. R. (2010). A Review of Uncertain Decision-Making Methods in Energy Management Using Text Mining and Data Analytics. doi:10.3390/en13153947
[13]. Jabbar, A., Iqbal, S., Tamimy, M.I. et al. Empirical evaluation and study of text stemming algorithms. Artif Intell Rev 53, 5559–5588 (2020). https://doi.org/10.1007/s10462-020-09828-3
[14]. Nurul, A. R., Muhammad, Z. Z., Sharifah, S. S.G., Noraini, S. (2021). Visualizing stemming techniques on online news articles text analytics. Bulletin of Electrical Engineering and Informatics. 10(1), pp. (365-373). ISSN: 2302-9285, DOI: 10.11591/eei.v10i1.2504
[15]. Alshanik, F., Apon, Amy A., Herzog, A., Safro, I., Sybrandt, J. (2020). Accelerating Text Mining Using Domain-Specific Stop Word Lists. Conference paper. https://www.researchgate.net/publication/346471844
[16]. Allahyari, M., Pouriyeh, S., & Assefi, M. (2017). Text summarization techniques a brief survey. arXiv:1707.02268
[17]. Abdi, A., Idris, N., Alguliyev, R. M., Alguliyev, R. M. (2016). An automated summarization assessment algorithm for identifying strategies. doi:10.1371/journal.pone.0145809
[18]. Xiao & Munro (2019). Text summarization of product tittles. In proceedings of SIGIR 2019. Workshop on eCommerce (SIGIR2019eCom), pp.1-7. Obtenido el 10 de febrero de 2021 desde: http://ceur-ws.org/Vol-2410/paper36.pdf.
[19]. Hans, C., Pramodana, A., M., & Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF) https://doi.org/10.21512/comtech.v7i4.3746.
[20]. Rahman, N., & Borah, B. (2018). Improvement of query-based text summarization using word sense disambiguation. Vol (6), pp (75-85). https://doi.org/10.1007/s40747-019-0115-2
[21]. Maridina, M., F., Portell, P. L., Boratto, L., & Fenu, G. (2020). A text mining approach to extract and rank innovation Insights from research projects. LNCS 12343, pp. 143–154. https://doi.org/10.1007/978-3-030-62008-0_10
[22]. Khan, A., Salim, N., Farman, H., Khan, M., Jan, B., Ahmad, A., Ahmed, I., & Paul A. (2018). Abstractive Text Summarization based on Improved Semantic Graph Approach. International journal of parallel programming. DOI: 10.1007/s10766-018-0560-3
[23]. Shivakumar, K., & Vishma, V. A. (2015). Text summarization using clustering technique and SVM technique. https://www.researchgate.net/publication/283831458
[24]. Larocca, N., J. Freitas, A. A., & Kaestner, C. A. A. (2002). Automatic Text Summarization using a MachineLearningApproach. https://www.cs.kent.ac.uk/people/staff/aaf/pub_papers.dir/SBIA-2002-Joel.pdf
[25]. Nur, A. F., Gulcin, O., & Cicekli, I. (2011). Text summarization using Latent semantic analysis. 37(4), pp (405-417). DOI: 10.1177/0165551511408848
[26]. Mohan, M. J. Shunitha, C. Ganesh, A. & Jaya, A. A study on ontology bases abstrative summarization. (2016). https://doi.org/10.1016/j.procs.2016.05.122
[27]. Yang, M., Li, C., Shen, Y., Wu, Q., Zhao, Z., & Chen, X. Hierarchical Human-like Deep neural networks for abstractive text summarization. doi: 10.1109/TNNLS.2020.3008037.
[28]. Gamma, E., Vlissides, J., Helm, R., & Johnson, R. (1995). Desing patterns. ISBN: 0201633612.
[29]. Morantach, N., & Gopalan, C. (2016). A survey on abstactive text summarization. International Conference on Circuit, Power and Computing Technologies (ICCPCT). pp.1-7. DOI: 10.1109/ICCPCT.2016.7530193