IJARSCT

ISSN (Online) 2581-9429

IJARSCT
Impact Factor: 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 1, November 2022

Synopser: Data Summarization Web Application

Prof. Pravin Latane[1], Sushant Bagul[2], Piyush Pate[3], Himanshu Patel[4], Sahil Pokharkar[5]
Professor, Department of Information Technology[1]
Students, Department of Information Technology [2,3,4,5]
Sinhgad Institute of Technology, Lonavala, Maharashtra, India

Abstract: *As there is an increase in the usage of digital applications, the availability of data generated has increased to a tremendous scale. Data is an important component in almost every domain where research and analysis are required to solve problems. It is available in a structured or unstructured format. Therefore, to get corresponding data as per the application purpose, easily and quickly from different sources of data on the internet, an online content summarizer is desired. Summarizers make it easier for users to understand the content without reading it completely. Abstractive Text Summarizer helps in defining the content by considering the important words and helps in creating summaries that are in a human-readable format. The main aim is to make summaries in such a way that they should not lose their context. Various Neural Network models are employed along with other machine translation models to bring about a concise summary generation. This paper aims to highlight and study the existing contemporary models for abstractive text summarization and also to explore areas for further research.*

Keywords: Data, Abstractive, Summarization, Neural Network.

I. INTRODUCTION

The amount of data available on the internet is increasing at an alarming rate. This data, on the internet, is unstructured. This unorganized form of data has made it difficult for a user to locate specific content. Eventually, the user spends time and resources to locate such relevant content. Hence, a content summarizer would be desirable to get a gist of an article and to check if the article answers the user's query before the user proceeds to read the article in detail. Automatic text Summarization is the elementary methodology used for any text summarization system. Automatic text summarizer has many applications in the industry: News Aggregators, Blogs, Product Descriptions, etc. The Text Summarization Technique paved the way for the development of summarization models. Initially, the text summarization was developed considering basic parameters and measures.
Text summarization can be classified as Abstractive and Extractive:

1.1 Extractive Summarization

Extractive summarization finds out the important sections of the content and creates a subset of sentences from the sentences present in the original document. It does not add new words to the existing content and cannot combine two or more sentences to compact the content. Based on the number of documents used as sources, summarization is further classified as single-sourced and multi-sourced. Extractive Summarization works based on combining the words or phrases from the corpus for the summary.

1.2 Abstractive Summarization

Abstractive summarization, on the other hand, analyses the whole content to reproduce the original content in a new and optimized way using advanced natural language techniques. The newly generated content is shorter and more importantly, conveys the most critical information of the original content. Abstractive summaries also generate fluent sentences that are grammatically correct, unlike extractive methods, which may lead to disfluent sentences.

II. LITERATURE SURVEY

Automatic text summarization techniques such as cluster-based, template-based, ontology-based, semantic graph-based methods, machine learning, fuzzy logic, and neural network approach have proven to be very useful over 50 years up to now. Automatic text summarization approaches are twofold extractive and abstractive summarization. Extractive

DOI: 10.48175/568

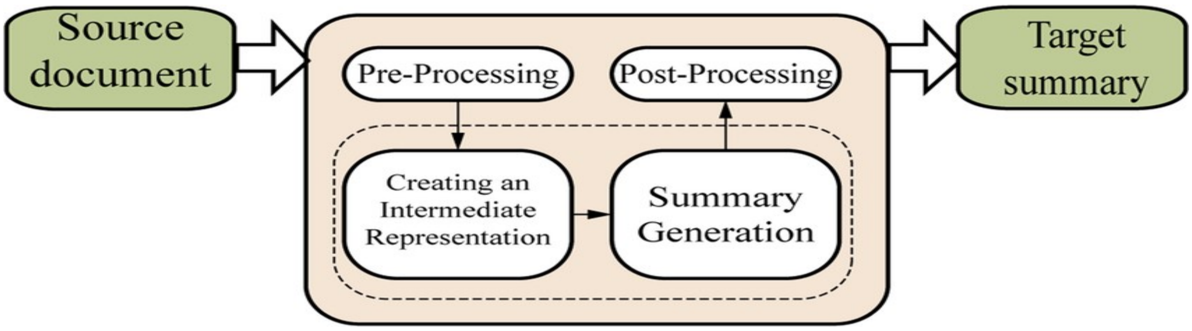
summarization techniques extract important sentences or phrases from the original text documents based on the weight of statistical and linguistic features and produce a summary without changing the original text. Abstractive summarization generates the summary by understanding the source text using linguistic methods and generating new sentences by improving the focus of a summary while reducing the redundancy rate. Abstractive summarization is more complex because it requires deeper analysis of source document(s) and concept-to-text generation. These techniques can be further classified as structured and semantic-based approaches. Semantic-based approaches provide more coherent, information-rich, and well-structured abstractive summaries than structured-based approaches. Due to the semantic representation, semantic graph-based approaches have been proven to be very successful over other methods for producing multi-document abstractive summaries.

Semantic-based methods are not able to achieve similar performance to deep learning approaches. Machine learning approaches to generate automatic summaries have greatly improved compared to the primitive text summarization methods, but they cannot be combined with the background information to obtain higher level abstraction and cannot scale to the requirement of large data sets. To understand a text well enough, we need some background knowledge. Some researchers were inspired by those ideas and used knowledge graphs to generate automatic text summaries. Knowledge graphs capture domain specifics and add rich and explicit semantics to infer additional knowledge and encapsulate a large amount of knowledge for human and machine consumption. Graph-based approaches use ranking functions comprising one or more sentence weighting features to determine the relevancy of each edge for inclusion in the summary.

According to the literature, we have identified that traditional word co-occurrence measures like TF-IDF give more importance to the words that are more frequent in the document. Hence, a pure ranking algorithm can be less effective because ranking alone cannot determine what kinds of features are selected for the summary. But, increasing the set of attributes to include semantic properties and topological graph properties in the ranking function yields statistically significant improvement for the abstractive summaries. Knowledge graphs lead to information overload, and hence proper summarization techniques need to be explored. Techniques in data mining are widely used to explore data from various perspectives and uncover unexpected relationships between pieces of text and summarize them into useful information. Therefore, computers use text mining to discover valuable and interesting information and knowledge techniques. Among these techniques, Association Rule from unstructured textual data in large volumes using different Mining (ARM) can be used in discovering the relationships for the decision-making process and discovering correlations between multidimensional features in the text documents.

III. IMPLEMENTATION DETAILS OF A MODULE

The source document is taken from the internet or from manually entering text paragraphs to get a summary. After receiving the text, that text is pre-processed. Pre-processing involves:



3.1 Removal of High-Frequency Words

This is done by comparing each word from the document to the list of high-frequency words. High-frequency words are those words that come a greater number of times in the text. The advantage of this process is not only the non-significance words are removed but also the size of the document can be reduced to 30 to 50%.

3.2 Suffix Stripping

In this step, each word is handled from the output of the first step. If any word is having a suffix, then the suffix gets removed and the word is converted to its original form.

3.3. Detecting Equivalent Stems

After suffix stripping, we have a list of words. Only one occurrence of the word is kept in the list. If two words have the same underlying steam then they refer to the same concept and should be indexed as such.

After that based on the keywords received, the system generates a new text summary that captures the most relevant information. And this summary is given to the user after post-processing of the text. In post-processing, similar meaning lines are removed.

IV. CONCLUSION

Machines will truly be intelligent when they summarize the text like humans. The approach to solving the Data Summarization problem using the abstractive method is really promising. This paper explains how we can use abstractive data summarization techniques to generate summaries.

REFERENCES

[1] 2020 @IEEE, A Survey on NLP-based Text Summarization for Summarizing Product Reviews by Ravali Boorugu, Dr. G. Ramesh

[2] 2020 @IEEE, Study on Abstractive Text Summarization Techniques by Parth Rajesh Dedhia, Hardik Pradeep Pachgade, Aditya Pradip Malani, Natasha Raul, Meghana Naik

[3] 2021 @IEEE, Multimedia Data Summarization Using Joint Integer Linear Programming by Sidhant Allawadi, Ritika, Vivek Rana, Minni Jain

[4] 2021 @IEEE, Techniques, and Research in text summarization- a Survey by Manish Shinde, Disha Mhatre, Gaurav Marwal

[5] 2020 @IEEE, Topic-Centric Unsupervised Multi-Document Summarization of Scientific and News Articles by Amanuel Alamo, Cori Lohstroh, Erik Madaus, Swati Padhee, Brandy Foster, Tanvi Banerjee, Krishnaprasad Thirunarayan, Michael Raymer

[6] 2020 @IEEE, Abstractive Web News Summarization Using Knowledge Graphs by M.V.P.T.Lakshika, H.A.Caldera, W.V.Welgama

[7] 2020 @Springler International Publishing, “An introductory survey on attention mechanisms in NLP problems,” in Intelligent Systems and Applications, by D. Hu, Y. Bi, R. Bhatia, and S. Kapoor

[8] 2020 @IEEE, “Dual encoding for abstractive text summarization,” by K. Yao, L. Zhang, D. Du, T. Luo, L. Tao, and Y. Wu

[9] 2019 @ International Journal of Scientific and Technology Research, “Best keyword set recommendations for building service-based systems”, by Dr. Gajula Ramesh, Dr.J.Somasekar, Dr. Karanam Madhavi, Dr. Gandikota Ramu,

[10] 2019 @Mind the Facts, “Knowledge-Boosted Coherent Abstraction Text Summarization” by Beliz Gunel, Chenguang Zhu, Michael Zeng, Xuedong Huang

[11] 2020 @PEGASUS, “Pre-Training with Extracts Gap-Sentences for Abstractive Summarization ” by Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu

[12] 2020 @ARXIV, “Text Summarization Techniques: A Brief Survey”, by Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trrippe, Juan B. Gutierrez, Krys Kochut.

BIOGRAPHY

- Prof. Pravin Latane:- Professor in the Department of Information Technology at Sinhgad Institute of Technology and contributed as a guide and author of the Project.
- Sushant Bagul:- An Undergraduate Scholar pursuing a Bachelor of Engineering in Information Technology from Sinhgad Institute of Technology. He is working under the guidance of Prof. Pravin. Latane.

DOI: 10.48175/568

- Piyush Pate:-An Undergraduate Scholar pursuing a Bachelor of Engineering in Information Technology from Sinhgad Institute of Technology. He is working under the guidance of Prof. Pravin Latane.
- Himanshu Patel:-An Undergraduate Scholar pursuing a Bachelor of Engineering in Information Technology from Sinhgad Institute of Technology. He is working under the guidance of Prof. Pravin Latane.
- Sahil Pokharkar:-An Undergraduate Scholar pursuing a Bachelor of Engineering in Information Technology from Sinhgad Institute of Technology. He is working under the guidance of Prof. Pravin Latane