Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020)
IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2

A Survey on NLP based Text Summarization for Summarizing Product Reviews

Ravali Boorugu
Assistant System Engineer
Tata Consultancy Services Limited
Hyderabad.
ravali.boorugu@gmail.com

Dr. G. Ramesh
Associate Professor
Department of CSE
GRIET, Hyderabad.
ramesh680@gmail.com

***Abstract*—No one can imagine life without a smartphone and internet nowadays. It has become essential for people of all age groups. With an increase in the usage of internet and smartphones, there has been a steady increase in online shopping too. Everyone wishes to get their products delivered at their home without any hassle. How to detect which products are genuine and pick the best among the unlimited options at the same price? Every user looks at the reviews before ordering anything online. Nevertheless reading those long reviews is not easy for everyone. Therefore, there must be something that can reduce the long reviews to short sentences of limited words depicting the same meaning. Text Summarization can come in hand in this aspect. Many NLP researchers are interested in Text Summarization. This paper is a survey on the various types of text summarization techniques starting from the basic to the advanced techniques. According to this survey, seq2seq model along with the LSTM and attention mechanism is used for increased accuracy.**

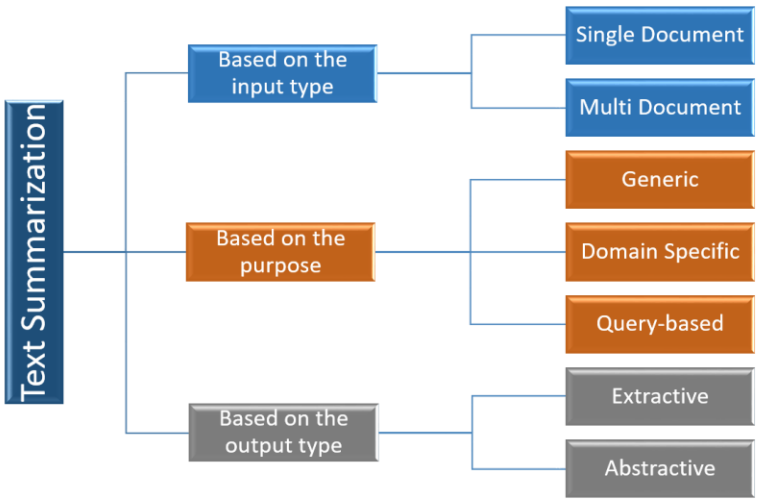
***Keywords*— Text Summarization, Product Review Summarization, NLP Techniques.**

I. INTRODUCTION

There has been a continuous increase in the number of internet users every year. With an increase in Internet users, comes a great deal of information that gets stored online every second. There is a need for summarizing this data without losing the original meaning of the data. Thus the process of Text Summarization comes into the picture with its benefits spread over different fields such as Machine Learning, Natural Language Processing, Artificial Learning, Semantics etc.,

Online Shopping has become a common thing these days as a wide variety of products are available at a single place. The ease of ordering a product and getting the product delivered directly to home at a convenient date and time has attracted many people. Along with these, the discount offers being offered by online shopping sites are making the people stick to online ordering. Everyone refers to the product reviews before buying a product. Then they can conclude which is the best product to buy among the different products available. Suppose a user needs to buy a laptop. Then he must go through different kinds of laptops available at his budget, make a note of different reviews for each product and choose the best among the available laptops. This is a tedious and time taking task. In addition to this, some users' reviews are so long that the user could get the actual meaning of it only after closely going through the review. Thus there is a need for minimizing the review to a shorter representative sentence which depicts the same meaning as the whole content.

This is the area where Text Summarization comes into picture with a great deal of benefits that could help us in choosing the best product from the whole lot. Text Summarization methods are broadly categorized into different types as shown in the figure below.



II. TEXT SUMMARIZATION TYPES

BASED ON INPUT TYPE: In this, had two types, Single Document and Multi-Document techniques.

Single Document Text Summarization (SDTS):

In this type of Summarization, the length of the input is short. There will be only a single document given as the input for Summarization. This was used in the early days of Text Summarization.

Multi-Document Text Summarization:

This is a process in which the length of the input on a particular topic is too long and therefore multiple documents are provided as an input for a summarization technique. This is often difficult when compared with the SDTS as there is a need to combine the summary of multi documents into a single document. The difficulty here is that there may be diversity in the themes of different documents. An ideal summarization technique often makes condenses the main themes maintaining readability, completeness and without missing the important sentences.

BASED ON OUTPUT TYPE: There are two types of techniques

Extractive Text Summarization:

As the name itself depicts, the extractive text summarization is the process in which the sentences are extracted from the whole text which could depict the similar meaning as the whole text but in a more condensed form. Majority of the text summarization techniques that are being used nowadays are of an extractive type.

Abstractive Text Summarization:

This is a more advanced type of text summarization which involves the formation of phrases or sentences that are not in the text but reflect the same meaning as the complete text. This method is more captivating but at the same time, it is more difficult for the model to form phrases or sentences that could bring the same meaning.

BASED ON PURPOSE: Here it is categorized into the below 3 types.

Generic Text Summarization:

The method in which the model makes no inferences about the meaning of the text to be summarized or any knowledge of the domain is called Generic Text Summarization. It makes a generic summary of the whole text, documents, photos or video clips.

Domain-Specific Text Summarization:

In this method of text summarization, the model uses knowledge of a specific domain like scientific documents, medical documents. This increases the accuracy and thereby gives a more meaningful, concise and easily understandable summary of the whole text.

Query-based Text Summarization:

This method involves taking a query as an input and based on that query, the model makes a summary of the text by selecting the sentences and phrases that are very much related to the query given as input.

III. LITERATURE SURVEY

There are many prominent works in Text Summarization from the past few years. Earlier works dealt mainly with Single Document Text Summarization. Now that the technology has increased as well as computing power has increased which paved the path for a faster, more effective and more accurate way of processing documents when compared with the earlier methods.

Niladri Chatterjee, Amol Mittal and Shubham Goyal in [5] proposed an extractive based Text Summarization technique that makes use of Genetic Algorithms. In this paper, they represented the single document as a Directed-Acyclic-Graph. Weight is given to each edge of the DAG-based on a schema explained in the paper. They use an Objective function to express the standard of the summary in terms such as ease of readability (readability factor), how closely sentences are related (cohesion factor) and topic relation factor. The Genetic Algorithm is intended to maximize the Objective function by selecting the prominent sentences from the whole text. Initially the Cohesion Factor i.e., how closely are the sentences related to each other is calculated. Then, the sentences that are similar to the input query should be given the highest preference called as Topic Relation Factor is calculated. After calculating the aforementioned factors, the Objective Function (fitness function) of the summary can be determined. Then a Genetic Algorithm is used to maximize the Objective function.

Amol Tandel et el in [6] proposed a multi-document summarization technique that will allow the customer to condense relevant data from multiple documents given as a single input. This method could save an ample amount of time along with increased efficiency. They have inspired from the then existing approaches like Cluster-based, Topic-based and Lexical Chain based. LexRank prevents the score maximization of Sentences that are not relevant to the main theme of the document. Lower scores are given to the sentences that contain noisy data because there will be no similitude with the cluster. In the initial phase, they will extract the summary of every single document. Then generated metadata from those documents. This metadata is used to construct a graph that shows how the sentences are relevant to each other by considering each document as a node and the appropriate weights are given based on the similitude of the metadata generated earlier.

Shivangi Modi & Rachana Oza in [11] discusses in detail about 3 single document techniques and 2 multi-document techniques.

Aditya Jain et el in [7] proposed a model which used Word Vector Embedding for Extractive Text Summarization. As per their paper, there are four prominent problems to deal with while extracting information. They are recognizing the most salient sentences from the document, removing the unnecessary information that is not relevant to the theme of the document, minimize the details and putting together the initially extracted information that is relevant into a condensed and organized report. To overcome the aforementioned challenges, they proposed a Word Vector Embedding approach to extract the prominent, then they used a Neural Network for Extractive Summarization by using Supervised Learning method. They tested on DUC2002 dataset and found that the results were more accurate when compared with the earlier summarizing methods. The results were satisfactory but can be improved if the size of the dataset is increased and theme diversity of the dataset and then implementing more effectual approaches like Sequence to Sequence Recurrent Neural Network for summarizing.

Nithin Raphal, Hemanta Duwarah and Philemon Daniel in [9] provided a review on the prominent research performed on the abstractive text summarization. As there are two methods of summarization: Extractive and Abstractive methods. The Extractive method as said by *Aditya Jain et el [7]* will select the prominent sentences that are in the document and make the summary out of it by maintaining the coherence between the sentences and sticking to the theme of the document. The Abstractive method, on the other hand, creates a summary by creating the phrases or sentences that may or may not be present in the document but could bring the complete meaning of the document. This is way more difficult than the extractive technique used earlier. It is very much similar to what a human could

generate after going through a document. Word embedding method and one hot vector methods failed to detect the similarly occurred word. This problem was resolved in *Mikolov et al* [1] [2] model in which they used continuous skip-gram-model, which takes input word and can project the probable contextual words whereas, on the contrary, the continuous bag-of-words model is exactly the converse of the CSG method.

They proposed various methods by which extractive summarization can be done, the preprocessing steps that are to be done in the initial phases, discussed the latest research in this arena, the various kinds of architectures, mechanisms involved, supervised and reinforced learning & the advantages and disadvantages of various architecture.

In [10] "Query-based Summarization using topic background knowledge" (2017) has been proposed. Basically, a query oriented approach means to develop the summary based on the query given as an input. As most of the queries don't hold the semantic details or information, the query-based model is not effective. So *Yang et al.* proposed a model that will use the search engines to develop background knowledge of the main theme of the document. Later they used the Page rank algorithm which contains the document information and cross-document information. They applied this algorithm on the document to construct the summary of the document. They used the China search engine Baidu for the building of theme background knowledge. In the future works that may be extended to Google, Yahoo etc. and the results can be compared with the earlier results. In this way can be built a more accurate summary as there is a good knowledge of the background theme of the document.

According to *Shi Ziyan in [3],* Summarization could not bring accurate results when the word has a lot of meanings. So there is a need for the particular domain knowledge of the main theme of the document as well. This brings the domain-specific text summarization into the limelight. But the problem arises when the referring is done inaccurately. Therefore this paper proposes a co-reference resolution algorithm to sort out this problem and bring accurate results. On the similar *lines Paul Gigioli, Nikhita Sagar, Anand Rao, Joseph Voyles [4] "Domain-Aware Abstractive Text Summarization for Medical Documents" (2018)* extended the domain-specific summarization by adding deep reinforced abstractive summarization method which is capable of going through the biomedical abstracts and summarizing them into a single line summary.

Priya Pawar et el in [12] discussed the importance of summarizing and classifying product reviews. They used hybrid classifiers such as SVM and Naïve Bayes. They also concluded that with the increase in the classifiers count the accuracy can also be increased.

Summarization of Online product review can be achieved with higher accuracy by using Seq2Seq model's, This model could bring a more accurate and very close summary of the product review submitted by the customer for a particular product.

IV. CONCLUSION

This paper explained in detail some of the remarkable works in the arena of text summarization. Summarization has always been a necessity for many years as there is a huge amount of information being released on the internet every day. This paper described all the major summarizations techniques and the prominent works that are being done on each technique. There has always been an improvement to the earlier technique which improved the accuracy like a single document summarization [5] has higher accuracy when compared with the multi-document summarization [6] and a domain-specific summarization [3] achieves higher accuracy when compared with the technique that has no prior knowledge of the domain. In similar lines, a query-based technique [10] yields the best results than the one without giving the query as input. From the above papers, could bring in the fact that the summarizers which are abstractive [4] [9] are way more efficient than the extractive [5] [7] ones but the former is more difficult than the latter one. As discussed earlier there are some remarkable works on product review summarization such as [12] which discussed the importance of summarizing online product reviews and the use of hybrid classifier i.e., SVM [13] and Naïve Bayes to develop a summary of the review. A seq2seq model is proposed for summarizing. Its advanced version Long Short Term Memory is used accompanied by attention-mechanism for getting a better accuracy of the summary developed.

V. REFERENCES

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean," Distributed Representations of Words and Phrases and their Compositionality," arXiv:1310.4546v1 [cs.CL], 2013.
[2] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781v3 [cs.CL], 2013.
[3] Shi Ziyan "The Design and Implementation of Domain-specific Text Summarization System based on Co-reference Resolution Algorithm" 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery.
[4] Paul Gigioli, Nikhita Sagar, Anand Rao, Joseph Voyles "Domain-Aware Abstractive Text Summarization for Medical Documents" published during 2018 IEEE BIBM.
[5] Niladri Chatterjee, Amol Mittal and Shubham Goyal's "Single Document Extractive Text Summarization Using Genetic Algorithms" (2012)
[6] Amol Tandel, Brijesh Modi, Priyasha Gupta, Shreya Wagle and Sujata Khedkar's "Multi-document text summarization - A survey" 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE).
[7] Aditya Jain, Divij Bhatia, Manish K Thakur's "Extractive Text Summarization using Word Vector Embedding"(2017).
[8] Canasai Kruengkrai and Chuleerat Jaruskulchai "Generic Text Summarization Using Local and Global Properties of Sentences" (2003).
[9] Nithin Raphal, Hemanta Duwarah and Philemon Daniel "Survey on Abstractive Text Summarization" 2018 International Conference on Communication and Signal Processing (ICCSP).
[10] Yang Wei and Yang Zhizhuo "Query based Summarization using topic background knowledge" 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)
[11] Shivangi Modi & Rachana Oza "Review on Abstractive Text Summarization Techniques (ATST) for single and multi documents" 2018 International Conference on Computing, Power and Communication Technologies.
[12] Priya Pawar, Siddhesha Tandel, Shweta Bore & Nikita Patil "Online Product Review Summarization" 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIEC).
[13] Dr. Gajula Ramesh, Dr.J.Somasekar, Dr. Karanam Madhavi, Dr. Gandikota Ramu, Best keyword set recommendations for building service-based systems International Journal of Scientific and Technology Research, volume 8, issue 10, October, 2019.