2023 International Conference for Advancement in Technology (ICONAT)
Goa, India. Jan 24-26, 2023

An Algorithmic Approach for Text Summarization

Amogh Joshi
Dept. of Computer Engineering
K. J. Somaiya Institute of Engineering and Information Technology,
Mumbai, India
amogh.dj@somaiya.edu

Prathamesh More
Dept. of Computer Engineering
K. J. Somaiya Institute of Engineering and Information Technology,
Mumbai, India
prathamesh.more@somaiya.edu

Soumya Shah
Dept. of Computer Engineering
K. J. Somaiya Institute of Engineering and Information Technology,
Mumbai, India
soumya.ms@somaiya.edu

Ms. Aarti Sahitya
Dept. of Computer Engineering
K. J. Somaiya Institute of Engineering and Information Technology,
Mumbai, India
aarti.sahitya@somaiya.edu

***Abstract*—With the advent of technology and increase in ease of access to digital devices, there has been a burgeoning increase in the flow and creation of information all across the internet. However, this has led to a lot of saturation and an increased amount of redundant information, which needs to be sorted through manually many times to find the important or pertinent information. But this ends up becoming a tiring and laborious task for many people. Hence, text summarizers, which are softwares that, when given a particular text input, summarize it effectively and output only the important parts in the text, while discarding the unimportant or redundant bits. A number of Text summarizers currently exist on the internet, for free and for a cost as well. Many of them are efficiently able to extract the important information and text out of the given input. However most of them are a single type of text summariser titled "Extractive Text Summarisers", which, while fairly accurate, is based on a model that merely extracts the important texts from the given input as it is, without specifically phrasing it in any other way, or without semantic understanding of the text itself. This leads to some inaccuracies in the summarized texts, such as some irrelevant words being put in the summarized output, even though they are unimportant, or some important sentences or phrases missing out on being in the output, as they have not been discussed enough in the input. That is why, the proposed system of Query based Summarizer not only works on queries, but is also aimed to be an Abstractive Text Summarizer, which will use semantic and syntactic understanding of the given input to summarize it and deliver a clear, concise and accurate output.**

Keywords—text, summarization, abstractive, extractive, paragraph, language, processing

I. INTRODUCTION

Newsletters, Articles and the World Wide Web are just a few of the many sources that provide a wealth of knowledge nowadays. For their social and professional needs, humans struggle with a lack of time while simultaneously wanting to find the most important information from a number of sources. This represents a query-oriented text summarizing method by extracting the most instructive sentences. To do this, numerous aspects from the sentences are chosen, and each one evaluates the importance of the phrases from a different perspective.[9]

The amount of information accessible online is expanding quickly. However, as a result of this, individuals want information abstraction or a summary more and more. In the era of information overload, text summaries have become a vital and practical tool for consumers to help them quickly understand the massive amount of information. Extractive and abstractive summarization are the two primary subcategories of text summarizing methods. Extractive summarization gathers significant sentences or phrases from the source materials to produce a summary without changing the original text. The abstractive summarization creates a generalized summary using sophisticated language creation and compression algorithms that effectively conveys information.[3]

A single-document summation (SDS), which creates the summary from a single document, or a multi-document summarization (MDS), which extracts the summary from a group of documents, are two options for the number of input documents that need to be summarized. Finding pertinent sentences from a single summary is a much easier procedure. Coherence is another challenge, and it's important to duplicate information across papers. There is a great deal of repetition because the distilled texts cover the same subjects.[8]

In an extraction-based summarizing method, a subset of words denoting the text's key ideas are extracted and concatenated to provide a summary. Imagine it as being similar to a highlighter that draws attention to the most important parts of a source text. In abstraction-based summarization, cutting-edge deep learning algorithms are used to simplify and paraphrase the original information in a way that is comparable to that of humans. Several techniques can be used for text pre-processing. The report highlights the most widely used methods, explains their benefits and drawbacks, and gives a brief overview of text pre-processing and feature extraction methods.

Depending on the task at hand, text summarizing methods can potentially be either extractive or abstractive. The most important sentences in the source documents were to be found and selected using an extractive technique, which involved leaving the sentences precisely as they were. Instead of just copying the most significant phrases from the book, abstractive summarization is a strategy for constructing a summary of a text based on its core concepts. Since it produces summaries that are similar to those of humans, abstractive summarization is more effective than extractive summarization, albeit it is more challenging to implement.[10]

978-1-6654-7517-4/23/$31.00 ©2023 IEEE

II. Literature Review

In the past twenty years, a number of researchers have conducted a variety of research on the topic of text summarization, particularly the summarization of English text. While still quite underdeveloped, in recent years, Text Summarization and research about it has been gaining popularity with the advent of newer techniques, more efficient pre-processing and state-of-art Machine Learning models which help in text summarization.

Particularly in the early 2000s, and the late 1990s, a number of researchers focused on devising newer techniques of text summarization, such as Kneser Ney smoothing, or Semantic text summarization. Rehman and Borah [1] used a common-sense knowledge based Semantic network to build a Query based text summarizer, wherein the Semantic relation score between the Query and the Input was calculated to decide if the sentence should be put into summary or excluded from it. Using a similar technique, Bellare et al. [2] used WordNet to identify relationships between various parts of input and query before traversing these relations breadth-wise to generate ranked Synsets. The rank then decided whether the sentence was included in the summary. Khan [3] conducted a literature review of various Text Summarization techniques, categorizing them by various approaches that they utilized. Kulkarni, Chammas, Zhu, Sha, Le [4] attempted to build a model to help automatically generate datasets for a Query-based Multi-document summarizer, or qMDS, which they titled 'AQUAMUSE'. Hogan [5] conducted a review of various Relation extraction or RE, its history, methods, limitations, and current methods. RE has four phases - pattern-based, statistical-based, neural-based, and large language model-based RE. Each of the phases was devised in order to overcome the shortcomings of the previous one. It presents various datasets and benchmarks used for evaluation of the RE performance based on some performance metrics. These data-sets focus on one of the two sub tasks; sentence-level RE and document-level RE.The model is evaluated using two techniques, corpus-based and instance-based. Kryscinski, McCann, Xiong, Socher [6] artificially generated datasets, which were then annotated. An uncased BERT model was fine tuned on the generated dataset. The model (termed as FactCC) was fed with source documents and claimed to generate a 2-way classification. Another version of FactCC (termed FactCCX) was trained with an additional span selection head using supervision of start and end indices. CNN/Dailymail dataset was experimented. W. El-kasaa, C. Salama, A. Rafea, H. Mohamed [7] conducted a comprehensive survey of all text summarization techniques, models, approaches and methodologies in use. Here, a survey of all existing text summarization techniques and types along with their classification was performed. An overview of techniques used as building blocks for automatic text summarization and their classification was provided. Ma, Zhang, Guo, Wang, Sheng [8] surveyed a number of deep learning based Multi-document Summarization (MDS) techniques, with works from 2015 to 2021 being collected. Nine different deep learning architecture design strategies, six deep learning based methods for text summarization were surveyed. Afsharizadeh, Ebrahimpour-Komleh, Bagheri [9] attempted to perform Query-oriented Text Summarization via the sentence extraction technique. Firstly, the data is prepared by applying preprocessing, then feature extraction from the sentences takes place. Then a score is assigned to each sentence based on its feature values. Sentences are ranked based on their scores. Top ranked sentences are selected to generate a summary. Abid [10] described a newer technique of Multi-Document Text Summarization using a deep-belief network. DBN and graphs are presented to implement MDS. The proposed model consists of four main stages. The first stage is concerned with applying preprocessing to every document in the dataset. A set of features is extracted in the second stage. The third stage is concerned with applying DBN to classify the sentences as important or unimportant. While the last stage is concerned with building a graph to select the most important sentences. Chen and Goodman [11] created novel techniques for analyzing the count-by-count performance of different smoothing techniques, including Modified Kneser-ney with three discounting values instead of one. James [12] examined a series of tests that were performed on variations of the modified Kneser–Ney smoothing model outlined in a study by Chen and Goodman. Two datasets were used: Heldout dataset, ATIS evaluation dataset. 1. Significance test proves the Chen and Goodman [10] choice of selecting parameter based on the extended contexts of n-gram is superior to selecting discounting parameter based on n-gram count 2.The discounting parameters for the smoothing model were calculated using the frequency of extended contexts, that is, the number of n-grams that have one, two, three, or four extended contexts. 3.Finally, the modkn–flex model yielded perplexities that were significantly lower than those for modkn–extend, but only for certain thresholds. Frydenlund, Singh, Rudzicz [13] offered a perspective on language modeling by considering it as a multi-label task and developing the methods necessary to train such a model. Further, they connected the idea of Label Smoothing to ranking via Knowledge Distillation by incorporating semantic similarity information from an ordered set of words. Ismail [14] extracted data from the Indonesian version of wikipedia, preprocessed it, fed those into SRILM toolkit, built n-gram lm with n=3,4,5; used kneser ney and wittenbell. Islam, Hossain and Arefin [15] performed a comparative analysis of different Text Summarization techniques using enhanced Tokenization. Four different approaches for Text Summarization were studied and explored. They are - Cosine Similarity based summarization, 2) Extractive summarization, 3) Text Rank based summarization, 4) Word count and heapQ based summarization. The four approaches were implemented on a singular 'Bangla' document and compared, with Text rank based summarization displaying the highest accuracy. Tabassum and Patil [16] conducted a survey on Text pre-processing and feature extraction techniques in natural language processing. The paper explored various text pre-processing techniques and the impact they can have on the accuracy of text-summarization and other NLP applications. Various pre-processing techniques such as NER, POS Tagging, Lemmatization etc are explored and studied, with their impact on training the model also being studied. Prabha and Parvathy [17] surveyed a number of applications of text processing and summarization as well as methodologies implemented for text summarization. A survey was made for various text summarization models and their performances were compared. Models such as Word Frequency, Cluster Based, Graph based were studied for Extractive Text summarization methods, while models such as Tree based, Rule Based, Ontology based were studied for Abstractive Text Summarization. It was noted that Abstractive text

summarization models generally yielded more accurate summaries as compared to Extractive Text Summarization models. Rudra, Goyal, Ganguly, Imran and Mitra [18] conducted research into using an extractive-abstractive approach for summarization of tweets during a crisis scenario. The use of Artificial Intelligence for Disaster Response (AIDR) for classification, summarization and tweet generation was proposed. The proposed system works by classifying tweets into categories via AIDR, then generating summaries based on scenario or disaster specific summaries. Finally, Kouris, Alexandridis and Stafylopatis used deep learning and semantic content generalization for generation of abstractive text summaries. The novel framework makes use of deep learning models of encoder-decoder architecture, as well as Abstractive text summarization using semantic-based (meaning-based) data transformation. The framework consists of three main parts - a Text generalization model, a deep learning network and a method for transforming the generalized summary into a humanly understandable and readable format. The model only generates a generalized summary, wherein rarer and less common words and phrases are replaced by more popular and commonly used words, which mean the same thing, which can then be summarized further and understood.

III. Research Gaps Identified

Extractive Methods of Text Summarization are inaccurate with a number of sentences just put into the summary based on the sentence ranking, which in-turn depends on the word frequency or phrase frequency. However, many times, words such as 'the', 'it', 'a', are far too common in a paragraph, and hence skew the rankings wrongly, resulting in unimportant sentences becoming part of the Summary, while important sentences get left out.

Abstractive Summarization techniques, albeit more accurate than Extractive summarization techniques in most cases, still have a few flaws, such as

- Intrinsic Factual Inconsistency Error - Contradictory facts make their way into summaries
- Extrinsic Factual Inconsistency Error - Neutral facts or random sentences with no relevance to input text make their way into summaries.

Finally, many Summarizers are also unable to take into account Context, and only able to summarize content based on the most common meaning associated with the word.

To overcome these errors and more, there needs to be a better system developed. Our proposed system which uses Kneser Ney Smoothing deals with the issue of summaries not being contextually accurate.

IV. Proposed System/ Algorithm

The system design is very simple to understand and to use, as its aim is to be as user-friendly as possible such that any user, who is unfamiliar with the use of digital electronic devices, can also use it with the same amount of ease as an accomplished user may. The system design consists of a simple user interface, which connects to the backend where the BART tokenizer is. The input given through the user interface input box is sent to the BART Tokenizer, which converts the input into tokens or 'tokenizes' the input, before forwarding it to the BART model. BART here stands for Bi-Directional Auto-Regressive Transformer, which is a

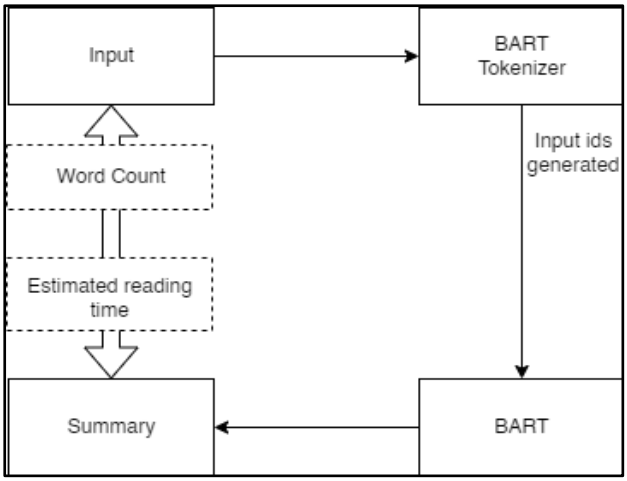

Fig. 1. Project Design

denoising autoencoder for pretraining sequence-to-sequence models. The model here has been trained using the 'facebook/bart-large' dataset. The proposed system provides an optimized way of generating query-based summarization.

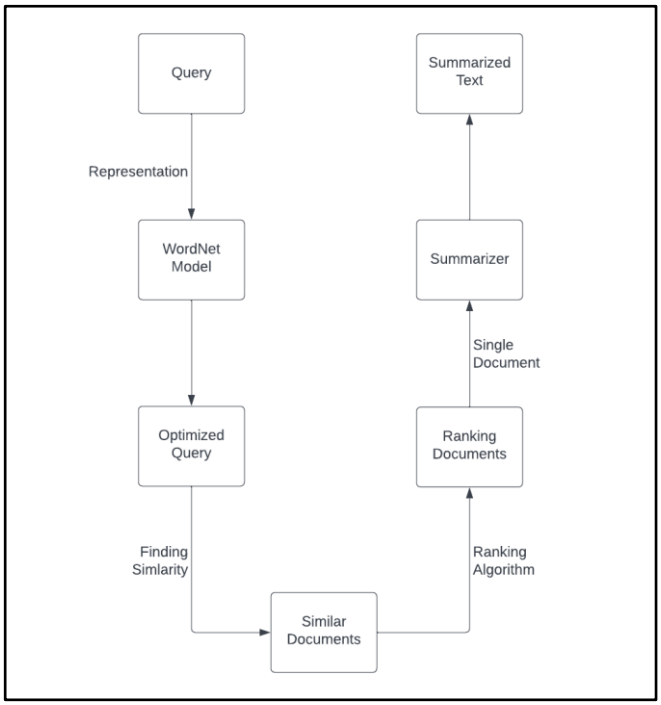

Fig. 2. DFD Level 1 Diagram

The queries entered by the user are first passed through an algorithm, which is context-based. The algorithm uses a word-net associated with the words in the query to process the query with context, thus finding out the true meaning behind the sentence and what the user wants. This optimized query is then searched in the documents. We get a list of documents having the answers for the query in it, which are then ranked using another algorithm, and then the highest-ranked document is selected. This document is then sent to a summarizer model to be summarized. For optimizing the query, we have proposed to use Kneser-Ney Smoothing technique to assign probabilities to the sentence. To assign scores to the sentences we have proposed to build a language model that will make use of the Kneser-Ney smoothing technique. So to train the language model we have proposed

to use the following system. The text is preprocessed and n-grams are found in that text. As the Kneser-Ney smoothing technique depends on the interpolation technique, discounts and backoff probabilities are calculated. Combining these probabilities we form an intuition, which is then used to score the sentence.[11]

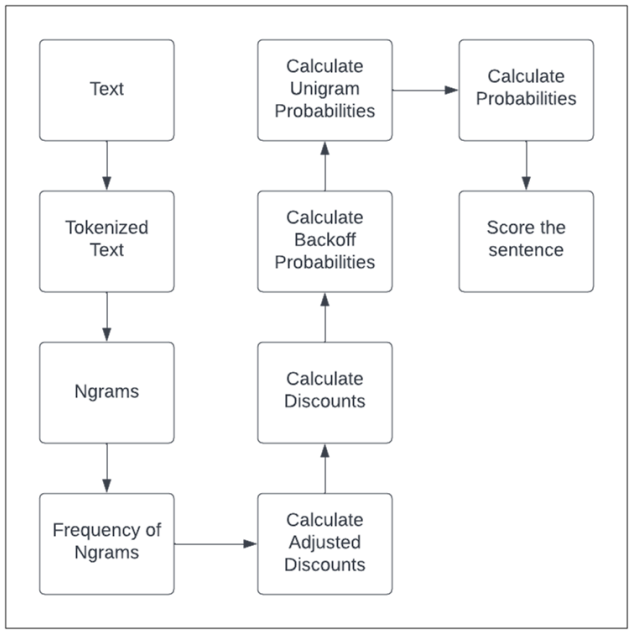

Fig. 3. Proposed Optimizer

A. *Figures and Tables*

Table 1 : List of Algorithms and Datasets. A number of algorithms were studied as part of the review, including the datasets they were tested on.

TABLE I. LIST OF DATASETS AND ALGORITHMS REVIEWED

Index	Survey		
	Name	***Dataset***	***Algorithm***
1	Improvement of query-based text summarization using word sense disambiguation	DUC 2005 and DUC 2006 Lesk Algorithm, Maximum Relatedness Score Jaccard Similarity for reducing redundancy	DUC 2005 and DUC 2006 Lesk Algorithm, Maximum Relatedness Score Jaccard Similarity for reducing redundancy
2	Generic Text Summarization using WordNet	DUC'2002 multi document summary corpus.	WordNet synset ranking, Cosine Similarity
3	AQUAMUSE: Automatically Generating Datasets for Query-Based Multi-Document Summarization	Colossal Clean Crawled Corpus	An automated approach to generate large datasets for the qMDS task for training and evaluating both abstractive and extractive approaches.
4	Evaluating the Factual Consistency of Abstractive Text Summarization	CNN/DailyMail dataset	Used a single-layer classifier model based on the [CLS] for checking factual consistency.
5	Query-oriented Text Summarization using Sentence Extraction Technique	DUC' 2007 Corpus,	Used Document Feature, Sentence Position, Topic and Topic Token Frequency, as well as Cluster, Start-Cluster and Bi-gram frequency among others to extend Ahuja method accuracies.
6	Multi Document Text Summarization using Deep Belief Network	DUC' 2004 Corpus.	Restricted Boltzmann Machine, Deep Belief Network, Jaccard Similarity
7	An empirical study of smoothing techniques for language modeling	WSJ/NAB Corpus, Brown Corpus	Additive smoothing, Jelinek-Mercer Smoothing, Katz Smoothing, Witten Bell and KneserNey smoothing
8	Modified Kneser-Ney Smoothing of n-gram Models	ATIS evaluation data corpus	Modified KneserNey algorithm
9	Language Modelling via Learning to Rank	Word-level Penn Treebank (PTB), WikiText-02 (Wiki02) datasets	Used and compared N-gram branching set construction and Plackett Luce Rank Loss
10	Comparison of Modified Kneser-Ney and Witten-Bell Smoothing Techniques in Statistical Language Model of Bahasa Indonesia	Extracted from wikipedia which is approximately 36M words in 1,7M sentences.	Markov Property, Modified KneserNey technique and WittenBell technique
11	Summarizing Situational Tweets in Crisis Scenarios: An Extractive-Abstractive Approach	Nepal Earthquake (NEQuake), Typhoon Hagupit/Ruby (Hagupit), Pakistan Flood (PFlood) Tweets and messages datasets	Used Content-words coverage vis-a-vis length, COWEXABS approach - unigram POS tagging based word-graph method for path generation, TF-IDF score & Informativeness score

Review Paper algorithms not covered

REFERENCES

[1] Rahman, N., Borah, B. Improvement of query-based text summarization using word sense disambiguation. Complex Intell. Syst. 6, 75–85 (2020). https://doi.org/10.1007/s40747-019-0115-2

[2] Kedar Bellare, Anish Das Sarma, Atish Das Sarma, Navneet Loiwal, Vaibhav Mehta, Ganesh Ramakrishnan, and Pushpak Bhattacharyya. 2004. Generic Text Summarization Using WordNet. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal. European Language Resources Association (ELRA).

[3] Khan, Atif. (2014). A Review on Abstractive Summarization Methods. Journal of Theoretical and Applied Information Technology. 59. 64-72.

[4] Kulkarni, S., Chammas, S., Zhu, W., Sha, F., & Ie, E. (2020). Aquamuse: Automatically generating datasets for query-based multi-document summarization. arXiv preprint arXiv:2010.12694.

[5] Hogan, W. (2022). An Overview of Distant Supervision for Relation Extraction with a Focus on Denoising and Pre-training Methods. arXiv preprint arXiv:2207.08286.

[6] Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2019). Evaluating the factual consistency of abstractive text summarization. arXiv preprint arXiv:1910.12840.

[7] "Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, Hoda K. Mohamed, Automatic text summarization: A comprehensive survey, Expert Systems with Applications, Volume 165, 2021, 113679, ISSN 0957-4174

[8] Ma, C., Zhang, W. E., Guo, M., Wang, H., & Sheng, Q. Z. (2020). Multi-document summarization via deep learning techniques: A survey. ACM Computing Surveys (CSUR).

[9] M. Afsharizadeh, H. Ebrahimpour-Komleh and A. Bagheri, "Query-oriented text summarization using sentence extraction technique," 2018 4th International Conference on Web Research (ICWR), 2018, pp. 128-132, doi: 10.1109/ICWR.2018.8387248.

[10] Azal Minshed Abid. (2022). Multi-Document Text Summarization Using Deep Belief Network. International Journal of Advances in Scientific Research and Engineering (IJASRE), ISSN:2454-8006, DOI: 10.31695/IJASRE, 8(8), 56–65. https://doi.org/10.31695/IJASRE.2022.8.8.7

[11] "Stanley F. Chen, Joshua Goodman, An empirical study of smoothing techniques for language modeling, Computer Speech & Language, Volume 13, Issue 4, 1999, Pages 359-394, ISSN 0885-2308

[12] James, Frankie. (2000). Modified Kneser-Ney Smoothing of n-gram Models.

[13] A. Frydenlund, G. Singh, and F. Rudzicz, "Language Modelling via Learning to Rank", AAAI, vol. 36, no. 10, pp. 10636-10644, Jun. 2022.

[14] Ismail, "Comparison of Modified Kneser-Ney and Witten-Bell smoothing techniques in statistical language model of Bahasa Indonesia," 2014 2nd International Conference on Information and Communication Technology (ICoICT), 2014, pp. 409-412, doi: 10.1109/ICoICT.2014.6914097.

[15] T. Islam, M. Hossain and M. F. Arefin, "Comparative Analysis of Different Text Summarization Techniques Using Enhanced Tokenization," 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2021, pp. 1-6, doi: 10.1109/STI53101.2021.9732589.

[16] Tabassum, Ayisha and Dr. Rajendra R. Patil. "A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing." (2020)

[17] PL. Prabha* and Dr. M. Parvathy, "Extractive and Abstractive Text Summarization Techniques," International Journal of Recent Technology and Engineering (IJRTE), vol. 9, no. 1. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 1040–1044, May 30, 2020 [Online].

[18] K. Rudra, S. Banerjee, N. Ganguly, P. Goyal, M. Imran, and P. Mitra, "Summarizing Situational Tweets in Crisis Scenario," Proceedings of the 27th ACM Conference on Hypertext and Social Media. ACM, Jul. 10, 2016 [Online].

[19] P. Kouris, G. Alexandridis, and A. Stafylopatis, "Abstractive Text Summarization Based on Deep Learning and Semantic Content Generalization," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1501