Original Article

Extractive - Abstractive Summarization Using Transformers: A Hybrid Approach

S. LourduMarie Sophie[1], Dr. S. Siva Sathya[2]

1 Research Scholar, Department of Computer Science, Pondicherry University, Pondicherry, India, [0000-0001-5345-598X],
2 Professor, Department of Computer Science, Pondicherry University, Pondicherry, India, [0000-0003-1009-6504]
1 lourdumariesophie15@pondiuni.ac.in
DOI: 10.47750/pnr.2022.13.S10.331

Abstract

Automatic text summarization (ATS) minimizes a lengthy text document into a condensed version by extracting the key points and main themes. Text summarising techniques can assist researchers in acquiring crucial information from various articles with less time and hassle. Several ATS systems have previously explored this area, and numerous text summarisation algorithms have been developed to extract essential details from textual sources and present them concisely. Unfortunately, many of these methods do not retain the text content's semantic elements and hidden meanings. Even though a substantial number of researchers have remedied these restrictions, they continue to provide a significant challenge for developing an effective article summarizer. This paper focuses on developing a framework for hybrid text summarization utilizing TF-IDF and Transformers to summarise research publications. Using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics, the effectiveness of the proposed hybrid summarizer for 10 research articles has been compared to the gold standard summary. Results demonstrate the similarity between the automated and human-generated summaries of the input article.

Keywords: Text summarization, Extractive approach, Abstractive approach, Term Frequency-Inverse Document Frequency (TF-IDF), Transformers.

I. Introduction

As more written content becomes available online, text summarization services have emerged as an effective and insightful resource for dealing with and making sense of this textual deluge. Everyone expects information to be short, concise, and automated in today's hectic environment. To give empowerment in the realm of computers, the notion of text summary receives the maximum emphasis in the digital revolution. Text summarization is a technology that improves information extraction methods and enables users to swiftly scan a huge collection of texts for relevant information [1]. In recent years, automated summarization has been acknowledged as one of the essential natural language processing (NLP) topics and one of the least resolved.

As illustrated in Figure1, each text summarising approach may be categorized according to its input, output, and purpose. Text summarization techniques can be either single-document or multi-document, depending on how many input documents are being analyzed concurrently [2], [3]. A summarizer can be extractive, abstractive, or hybrid based on its outcome. An extractive summarizer determines the most significant terms in the input document, retrieves those aspects, compiles them, and finally creates the summary [4]. In contrast, an abstractive summarizer detects the most significant concepts and provides a new text that represents those concepts in a different way [5]. Depending on its intended usage, a summarization technique might be either generic or query-based. A generic summary provides simply the basic concepts of the input, but a query-based summary includes information retrieved in accordance with the user's needs, which are provided to the system through a query or key phrase [6]. The methodology for the summary that is detailed in this article is extractive, single-document, and generic.

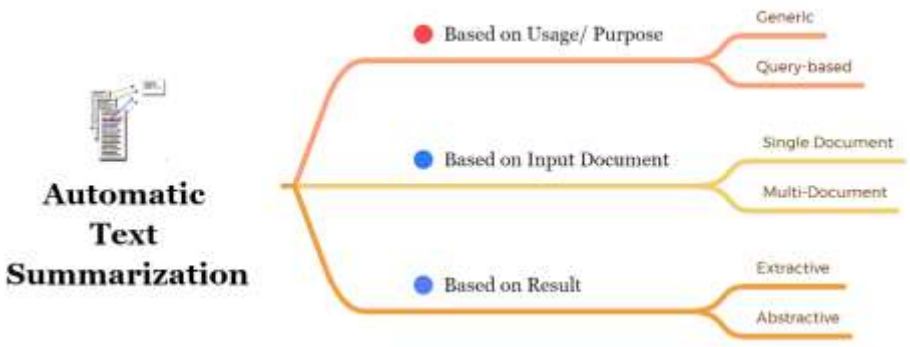

Figure 1: ATS Classification

Although extractive approaches may be suitable for finding the most pertinent data, they still lack the clarity and coherence of summaries created by humans. Abstractive summarization approach produces a few lines that effectively distil the main points of a given text material. Abstractive summaries have proven to be the most potential in resolving challenges associated with retrieving critical details from textual data. However, an abstract summary may generate phrases not present in the original source. The attention-based encoder-decoder strategy in the abstractive approach has lately been extensively investigated, in part due to the effectiveness of neural networks in machine translation studies [7]. Current abstractive summarization methods' effectiveness is insufficient to warrant widespread adoption [8].

This research presents a hybrid article summarizer that combines extractive and abstractive approaches to produce the final summary. The extractive method employs the TF-IDF approach to generate an extractive summary, which is then passed as input to the Text-to-Text Transfer Transformer (T5) [9] to produce an abstractive summary. The proposed summarizer solves a significant difficulty in summarising, namely optimally compressing the source content while keeping its core notions.

The rest of the paper is arranged in the following order: Section II provides the relevant works. Sections III and IV offer the proposed methodology and experimental results. Section V expresses concluding observations and outlines future research directions.

II. Related Works

The earliest study on automated summarising began in the 1960s with the introduction of "Edmundson Paradigms" [10], a structure for extractive summarization. Subsequently, several researchers concentrated on this issue and endeavoured to enhance the effectiveness of summaries by proposing novel methods and data sets. Due to the restricted processing power, the findings were not particularly impressive. The most significant advancements in this field have occurred in the past two decades, where academics embracing statistical and computational approaches have presented a variety of summarization tools to gain access to sophisticated computing resources [11].

Numerous approaches for summarising have been developed in recent times. These methods incorporate the data supplied by a frequent itemset approach (like TF-IDF statistics) to choose the phrases that are the most representational of the whole document and do not include duplicate information [12].

A generative pre-trained transformer (GPT) [13] trained for diverse NLP applications was built in 2018. On the basis of this strategy, a bidirectional encoder representation for transformers (BERT) employing language models and word embeddings was proposed to understand deep bidirectional interpretations [14]. BERT is complicated in the context of computing but far more accurate practically. One of the shortcomings of this approach is its inability to recognize and comprehend negation.

Reference [9] presents a paradigm where each NLP process is viewed as a text-to-text challenge. In other words, the input texts are used to make output messages. The T5 (text-to-text transfer transformer) model is designed for transfer learning on NLP applications utilizing recurrent neural networks (RNNs) and a transformer framework

with a self-attention feature. Self-attention is a form of attention that analyses a series of phrases by substituting each element with a weighted mean of the remaining items. T5 is trained using the C4 (Colossal Clean Crawled Corpus) dataset. Large-scale tests using the T5 model have shown that they can enhance self-supervised learning in natural language processing. In this context, [7] created a multi-sentence abstract using the T5 model; experiments on News dataset demonstrated that the T5 excels at abstractive summarization.

III. Proposed Methodology

In this work, an extractive summary is initially constructed employing the Term Frequency-Inverse Document Frequency approach, keeping the document's themes in mind, to build abstractive summaries for research papers. The abstractive summary is produced by sending the segmented extractive summary to the T5 model. The proposed framework is illustrated in Figure 2.

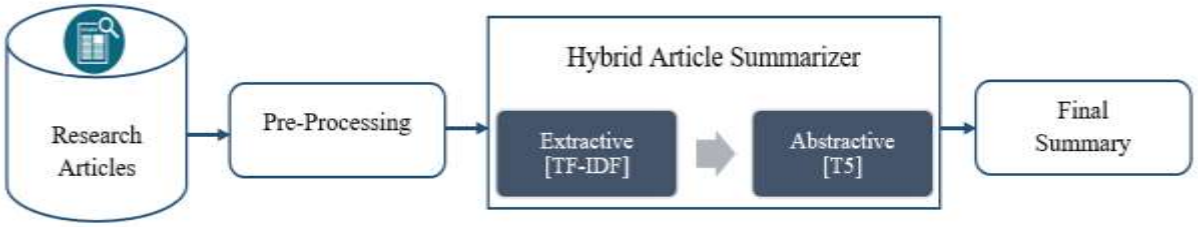

Figure 2: Proposed framework

3.1 Extractive summarization approach

The extractive method finds the most important passages in each document, integrates them into a holistic unit, and then delivers them to the reader. For this work, a TF-IDF approach is utilized to generate an extractive summary for the input research article. It is a statistical method for extracting useful information from large amounts of data by comparing the frequency of terms in one document to the inverse proportional frequency of those words in another set of documents. It implies that the user may attribute greater significance to words often appearing throughout a manuscript [15]. However, if that identical term appears in several other texts, it loses any significance. The calculation of term frequency and inverse document frequency is given in equations (1) and (2), respectively,

$$\mathrm{tf(t)} = \frac{\text{Total number of times term t appreared in A}}{\text{Total Number of terms in A}} \qquad (1)$$

$$\mathrm{idf(t)} = \log \frac{\text{Count of } A_n}{\text{Total count of } A_n \text{ with term t in it}} \qquad (2)$$

Here, An denotes a specific article, An denotes a set of articles. TF-IDF for a term t in the article is computed as given in equation (3),

$$\mathrm{tf - idf\,(t)} = \mathrm{tf(t)} * \mathrm{idf\,(t)} \qquad (3)$$

Once the input is pre-processed, a frequency value is computed for each word and verb in the article. At this point, the score for each sentence is calculated using the TF-IDF approach. The sentences with the highest relevance ratings are chosen and then pieced together to form the extractive summary.

3.2 Abstractive summarization approach

In this approach, data extracted from articles are fed into a language generation machine, which produces a summary with phrases that differ from the original text. A text-to-text model is employed for this work to generate an abstractive summary from the extractive summary.

In the realm of Transformer-based natural language models, T5 ranks among the most cutting-edge. It employs numerous pre-training objectives, like classification, translation, and summarization, where each target is portrayed as a language generation activity. The T5 framework uses relative positional embeddings instead of sinusoidal encoding [16], and it includes the same number of encoder and decoder blocks as the previously intended Transformer design. Tokens in an input sequence are embedded into a sequence of vectors and then sent to the encoder in this paradigm. The encoder is composed of a series of "blocks," and each block has a self-attention layer and a rudimentary feed-forward network [9]. In general, the system takes in one text and gives out another. During the process of the development of the output sequence, the model returns to the input and focuses on the data that is significant to it by performing dynamic access to the pertinent bits of data relying on the hidden units of the decoder. In the case of the summarization task, in this stage, the T5 model is fed the extracted summary partitions, and it generates a paragraph that serves as an abstract summary for each partition.

IV. Experimental Evaluation

4.1 Dataset

As far as we are aware, there isn't a specially-created dataset for summarising text. Therefore, researchers have relied heavily on research articles and their abstracts to provide concise summaries of their findings. 10 articles were arbitrarily chosen from the Science Direct database [17] and used to analyze and assess the performance of the proposed hybrid summarizer. The complete texts of these publications have been fed into the hybrid summarizer, with the abstracts serving as a benchmark against which the proposed solution's abstractive summaries may be evaluated. On an average, each article has about 3000 sentences and the summary contained approximately 700 sentences.

4.2 Data Pre-processing

Pre-processing the articles involves eliminating punctuation, prepositional phrases, parts of speech, changing words to lowercase, deleting stopwords, and identifying the stemmed words. The process of stopword removal involved excluding all of the terms found in the English stopword dictionary. Once stopwords are eradicated, each word in the document is reduced to its root form. Python programming language was used for implementation purposes.

4.3 T5 model hyperparameter tuning

The manual configuration strategy was used to set the T5 specifications, where the training batch size and epochs are set to 2. In light of the fact that we are currently developing this hybrid model, we have selected only 10 articles for processing in this work. Sample functions from the pandas framework are used to divide the dataset into a training set (70%) and a testing set (30%). Corresponding precision and loss values for each input are calculated and examined. During the inference stage, which follows the training step, the model is fed with the testing data to generate the abstract summary.

4.4 ROUGE metric

Researchers frequently utilize ROUGE (Recall Oriented Understudy for Gisting Evaluation), a benchmark metric for automated summary assessment, to evaluate the efficacy of generated summaries. When comparing a gold-standard (human) summary to a system-generated summary, ROUGE scores are based on the percentage of shared terms between the two. In this case, the ROUGE-N variant is used, but there are many more. ROUGE-N gives the proportion of n-grams shared between the system and human summary. To evaluate performance using the ROUGE measure, we compare the output summary to the abstract of each article. The findings are tabulated in Table I, and their visual illustration is given in Figure 3.

Table I. Average performance measure of 10 research articles

ROUGE Metric	Average Performance Measure of Proposed Summarizer
Precision	0.801
Recall	0.744
F1 score	0.766

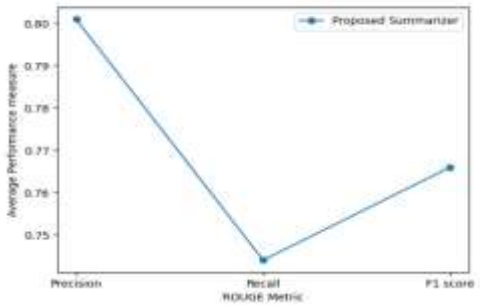

Figure 3: Average performance measure for 10 articles

V. Conclusion and Future work

In recent times, various text summarising strategies have been created in order to derive significant information from textual materials and display it in a condensed version. These methods' prime limitation lies in identifying the concepts that effectively express the text's core issue and extracting the sentences that best describe these vital notions. The second limitation is the right interpretation of the fundamental concepts in order to develop new paraphrased phrases that are not identical to the original text. Both limitations are considered in the proposed work while creating the hybrid summarizer. This work presents a circuitous approach for producing abstract summaries for research articles by employing frequency and transformer-based techniques. Experiments were conducted on 10 articles taken from Science direct database. In future, we plan to extend this work to a larger dataset. Also the proposed summarizer lacks effectiveness, this can be overcome by fine-tuning the models' parameters. Another future research direction includes enhancing the proposed summarizer to perform summarization on multiple documents.

References

[1] A. P. Widyassari et al., "Review of automatic text summarization techniques & methods," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 4, pp. 1029–1046, Apr. 2022, doi: 10.1016/J.JKSUCI.2020.05.006.

[2] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," Expert Syst. Appl., vol. 165, Mar. 2021, doi: 10.1016/J.ESWA.2020.113679.

[3] C. Shah and A. Jivani, "Literature study on multi-document text summarization techniques," Commun. Comput. Inf. Sci., vol. 628 CCIS, pp. 442–451, 2016, doi: 10.1007/978-981-10-3433-6_53.

[4] A. Sahoo, D. Kumar Nayak, and M. Tech Student, "Review Paper on Extractive Text Summarization," Int. J. Eng. Res. Comput. Sci. Eng., vol. 5, pp. 2394–2320, 2018.

[5] M. A. I. Talukder, S. Abujar, A. K. M. Masum, S. Akter, and S. A. Hossain, "Comparative Study on Abstractive Text Summarization," 2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020, Jul. 2020, doi: 10.1109/ICCCNT49239.2020.9225657.

[6] A. Aker, L. Plaza, E. Lloret, and R. Gaizauskas, "Multi-Document Summarization Techniques for Generating Image Descriptions: A Comparative Analysis," pp. 299–320, 2013, doi: 10.1007/978-3-642-28569-1_14.

[7] E. Zolotareva, T. M. Tashu, and T. Horváth, "Abstractive text summarization using transfer learning," CEUR Workshop Proc., vol. 2718, no. August 2020, pp. 75–80, 2020.

[8] S. Alhojely and J. Kalita, "Recent Progress on Text Summarization," 2020 Int. Conf. Comput. Sci. Comput. Intell., pp. 1503–1509, Dec. 2020, doi: 10.1109/CSCI51800.2020.00278.

[9] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1–67, 2020, Accessed: Nov. 10, 2022. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[10] H. P. Edmundson, "New Methods in Automatic Extracting," J. ACM, vol. 16, no. 2, pp. 264–285, Apr. 1969, doi: 10.1145/321510.321519.

[11] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," undefined, vol. 47, no. 1, pp. 1–66, Jan. 2016, doi: 10.1007/S10462-016-9475-9.

[12] E. Baralis, L. Cagliero, S. Jabeen, and A. Fiori, "Multi-document summarization exploiting frequent itemsets," undefined, pp. 782–786, 2012, doi: 10.1145/2245276.2245427.

[13] Q. Zhu and J. Luo, "Generative Pre-Trained Transformer for Design Concept Generation: An Exploration," Nov. 2021, Accessed: Nov. 11, 2022. [Online]. Available: http://arxiv.org/abs/2111.08489.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 4171–4186, Oct. 2018, doi: 10.48550/arxiv.1810.04805.

[15] L. Yao, Z. Pengzhou, and Z. Chi, "Research on News Keyword Extraction Technology Based on TF-IDF and TextRank," undefined,

pp. 452–455, Jun. 2019, doi: 10.1109/ICIS46139.2019.8940293.
[16] T. R. Goodwin, M. E. Savery, and D. Demner-Fushman, "Towards zero-shot conditional summarization with adaptive multi-task fine-tuning," Find. Assoc. Comput. Linguist. Find. ACL EMNLP 2020, pp. 3215–3226, 2020, doi: 10.18653/V1/2020.FINDINGS-EMNLP.289.
[17] "ScienceDirect." https://www.sciencedirect.com/.