Extractive Text Summarization of Research Papers in Health Sector

Aditi Joshi
dept. of Computer Science and Engineering
The Ohio State University
Columbus, USA
joshi.463@osu.edu

Harshini Kavuru
dept. of Computer Science and Engineering
The Ohio State University
Columbus, USA
kavuru.7@osu.edu

I. Introduction

Reading and understanding through numerous research papers is integral to starting one's research work. Unfortunately, information overloading has become a widespread problem faced while exploring for information on any particular topic. Moreover, reading various articles may become time-consuming and monotonous. Abridging this information will enable experimenters to expand their scope of investigation in the available time. Text summarization is an extensively used technique for reducing the content while conserving the essential parts of it. Recently, there has been plenty of research in health sector incorporating AI techniques. Consequently, it has become difficult to study the extensive existing work in this area. In this project, we aim to develop a comprehensive text summarizer that summarizes text and enables researchers to get precise summary for their work.

II. Dataset

A. Data Collection

As a part of the data collection step, we will be collecting research articles from well established databases like Academic Search Complete, Scopus, IEEE Xplore, Web of Science, PubMed, Springer, etc.

As an example please find the link for research articles below-

Database Link

III. Main hypothesis

The project's main objective is to create an efficient text summarizer that will generate concise text summary from the input research paper. In this project we would be focusing on research papers in health domain which utilizes different AI techniques. The generated summary will consists of all the essential parts of the article, like the title, objective of the research, technology and proposed architecture, methodology, datasets, evaluation parameters and results. It will give the reader a precise idea about the entire paper. This will help the researcher to decide if the current research paper is relevant for their study or not.

Although numerous approaches are available for pre-processing and processing the given text, choosing the most promising approach for the provided text is challenging. Hence, we propose an AI methodology for this project to produce the best possible text summary for the given research paper. We aim to optimize our natural language processor with the best possible algorithm and deliver an outcome that furnishes the reader with all the vital information in the given research paper. After producing the outcome, the Rouge-1, Rouge-2, and F-1 score metrics will be employed as the evaluation metrics.

Literature Review

In this paper [1], the input texts are read sentence-wise, and weights are assigned to them. The sentences are then ranked in order of their weight. The sentences with high weight are given high priority and are presented in the summary. Finally, the summary will be converted into an audio file and given as the output to the users. Here, a novel statistical approach is implemented to extract text summary from a single document. The obtained outputs are evaluated by comparing them to the human summarized data. For the results, five documents are tested with the proposed methodology. Each of these papers has 20 sentences. Only those sentences whose rank is higher than eight are considered in summary. [1]. In this paper [2], the authors not only summarized the central idea and text but also considered the image and table captions. The Recurrent Neural Network (RNN) and BiLSTM are implemented in the proposed methodology. In addition, the "non-anonymized version of CNN and Daily Mail" are used as the dataset. The results show that the proposed model outperforms the unsupervised "lead-3 method, and the REFRESH model". [2]. The paper focuses [3] on summarization depending on sentence similarity measures along with sentence features. It employs similarity measures between sentences, Estimating the number of clusters, Sentences Clustering, and Topic Sentences Extraction. DUC2003 - document dataset is used for the project. ROUGE-1, ROUGE-2, and F1-Measure are used to establish that the proposed method is better than other summarization methods [3]. In this paper [4], a greedy extractive summarization technique is used to summarize scientific papers from the "arXive and PubMed" database. Here, "Greedy Extractive Summarization" algorithm along with "Variable Neighborhood Search (VNS)" is employed in the proposed method. The study demonstrated that more

work could be done on extractive summarization techniques since the best ROUGE-1 score is typically 0.55, and the most advanced models cannot surpass a 0.50 level. [4].The paper [5] performs two-stage summarization using extractive and abstractive techniques. It employs a sentence similarity matrix and beams search algorithm for extracting the text. Chinese data - 1200 news texts (portals Sina, NetEase) and DUC-2004 are used as datasets. The proposed model improved by 3% on DUC-2004 and an improvement of + 2.56 ROUGE1, + 1.33 ROUGE2, and + 1.53 ROUGEL for the Chinese dataset. [5].The paper [6] uses unsupervised learning for text summarization, reducing the need for huge datasets to train the model. The "Learning Free Integer Programming Summarizer (LFIP-SUM)" is used for enforcing unsupervised learning on the document. In addition, "the non-anonymized version of CNN/Daily Mail, the Wikihow, and Cornell Newsroom" dataset is used as datasets. The suggested architecture revealed that 96.6% of the entire number of documents are summarized in fewer than three sentences. Furthermore, 52.5% and 32.6% of the content are summed in two and one sentences, respectively. [6].The purpose of [7] paper was to present a survey on the latest deep learning techniques used in text summarization. Here, the authors have presented a large variety of datasets and evaluation measures that can be used while dealing with extractive text summary. The author has discussed following algorithms -Deep Learning Techniques: "Restricted Boltzmann Machine (RBM), Variation Auto-encoder and Recurrent Neural Network (GRU, LSTM)". In the survey research paper, the author has highlighted various data bases which can be used to extractive text summary, namely – "1) Daily Mail (Number of documents: Training -196,557 Validation -12,147 Testing- 10,396), 2) DUC2006 (50 corpuses), 3) DUC2002 (576 documents and summary), 4) DUC2007 (45 corpuses), 5) Summarization and Keyword extraction from emails (SKE) - emails 6) British Columbia University (BC3) - emails 7) Essex Arabic Summaries Corpus (EASC)" has 153 Arabic documents. The discussed techniques were evaluated on the basis of ROGUE metric [7].

CHALLENGES

Some of the challenges we would face while completing the project are:

1) **Collecting Data:** Even though there is plenty of research available in health sector domain, collecting an good quality database with appropriate research papers will be challenging.

2) **Methodology:** The project's primary goal is to provide the best possible summary for the given paper. The major role in achieving the goal depends on the methodology implemented. There are a lot of approaches available for text summarizing, and there is also the possibility of formulating a novel technique for solving the problem. Determining the most promising technique for the collected data will be a challenging task.

3) **Range of papers:** The increasing scope of research has improved the range of research papers available. Each paper has a novel approach implemented for addressing the problem statement; consequently, it will be a challenge for our proposed system to highlight novelty of the input research paper.

4) **Contextual words and phrases:** The same word can have a different meaning in different contexts. Hence, it becomes difficult for the machine to understand and comprehend the words and phrases based on the sentence. Furthermore, the word that has the highest significance in one paper might not be noteworthy in another.

ANTICIPATED RESULTS

1) The summarized text should be in the given word limit.

2) The proposed algorithm should be able to successfully extract essential information from the input research paper.

3) The implemented algorithm should outperform the already existing algorithms. This will be validated with the help of evaluation parameters such as ROGUE and F1 score.

REFERENCES

[1] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

[2] L. Chen and M. L. Nguyen, "Sentence Selective Neural Extractive Summarization with Reinforcement Learning," 2019 11th International Conference on Knowledge and Systems Engineering (KSE), 2019, pp. 1-5, doi: 10.1109/KSE.2019.8919490.

[3] P. -y. Zhang and C. -h. Li, "Automatic text summarization based on sentences clustering and extraction," 2009 2nd IEEE International Conference on Computer Science and Information Technology, 2009, pp. 167-170, doi: 10.1109/ICCSIT.2009.5234971.

[4] I. Akhmetov, A. Gelbukh and R. Mussabayev, "Greedy Optimization Method for Extractive Summarization of Scientific Articles," in IEEE Access, vol. 9, pp. 168141-168153, 2021, doi: 10.1109/ACCESS.2021.3136302.

[5] W. Liu, Y. Gao, J. Li and Y. Yang, "A Combined Extractive With Abstractive Model for Summarization," in IEEE Access, vol. 9, pp. 43970-43980, 2021, doi: 10.1109/ACCESS.2021.3066484.

[6] M. Jang and P. Kang, "Learning-Free Unsupervised Extractive Summarization Model," in IEEE Access, vol. 9, pp. 14358-14368, 2021, doi: 10.1109/ACCESS.2021.3051237.

[7] D. Suleiman and A. A. Awajan, "Deep Learning Based Extractive Text Summarization: Approaches, Datasets and Evaluation Measures," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2019, pp. 204-210, doi: 10.1109/SNAMS.2019.8931813.