Abstract Generation with Hybrid Model Supported by Relevance Matrix

Student - R.P.D. Kumarasinghe (219354V)

Supervisor - Dr. Nisansa de Silva

Overview

Overview

1. Introduction
2. Research Problem
3. Research Objectives
4. Related Work
5. Dataset
6. Methodology
7. Results
8. Conclusion
9. Publications
10. References

1. Introduction

Introduction

A research paper is a combination of various sections. Commonly used sections identified are as follows

- Abstract
- Introduction
- Related Work
- Methodology
- Experiment
- Results
- Conclusion

Introduction

Abstraction Section

The abstract of a research paper provides a quick summary of the entire paper from problem to solution to the results described in subsequent sections.

Abstract is meant to be

- Concise
- Informative

  Abstract often dictates whether a reader will invest time in reading the full document, making it a vital part of academic writing

2. Research Problem

Research Problem

1. Alleviate the burden on researchers by developing a method for automatically generating abstract sections based on the content of the subsequent sections of the paper

2. Enhance the overall quality and consistency of scientific abstracts across disciplines. By allowing researchers to focus on fine-tuning the generated abstracts rather than creating them from scratch

3. Research Objectives

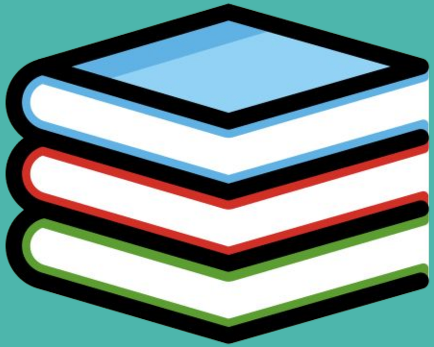
Research Objectives

1. Creating a Sufficient Dataset: Develop a comprehensive dataset tailored for the task of abstract generation. This dataset will be section-wise broken down to facilitate detailed analysis and model training.

2. Evaluating Existing Summarization Technologies: Assess the performance of state-of-the-art text summarization technologies on the newly created dataset as well as other comparable datasets.

3. Developing Automatic Summarization Models: Create advanced summarization models capable of generating abstracts automatically.

4. Related Work



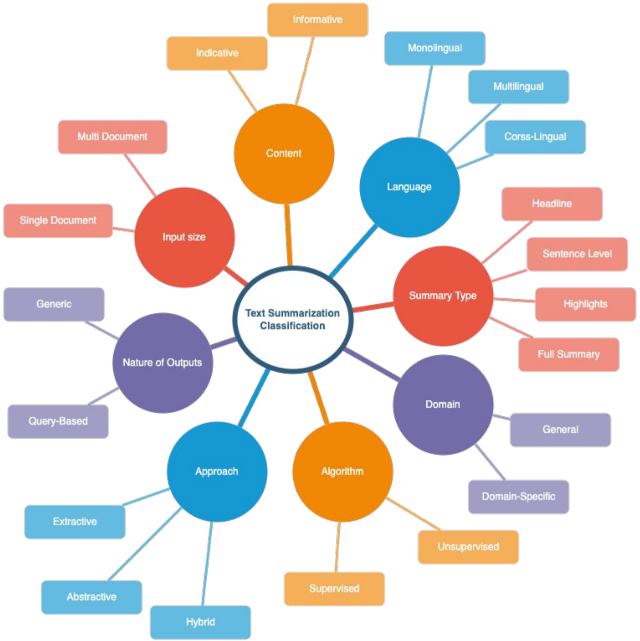
Related Work : Popular Tools

- **Resoomer [1]** application creates accurate text summaries, allowing users to quickly scan publications for key subjects, find essential facts and ideas, and comprehend articles. Long content can be summarized in just 500 words for users
- **SummarizeBot [2],** With the summarization of extensive texts, an AI and blockchain powered solution allows users to learn more while reading less. But it is a general-purpose tool and may not perform as well with highly specialized or technical content
- **SMMRY [3]** is programmed to condense a TXT or PDF text into the most significant sentences in just 7 seconds, making it a quick and efficient approach to grasp information. It is used by Reddit's AutoTLDR bot to generate short summaries of long Reddit submissions
- **TLDRThis [4]** is an intuitive online tool and browser extension designed to generate concise summaries of articles and lengthy texts.
- **TextCompactor [5]** also has a language translation function. Furthermore, it enables the option of several versions of these summarizations with the 'summarization ratio' in addition to generating text summaries. Users can vary the density of the paraphrase by changing the percentage from 5% to 80%.

[1] Resoomer, "Summarizer to make an automatic text summary online." [Online]. Available: https://resoomer.com/en
[2] SummarizeBot, "Get to know more by reading less!" [Online]. Available: https://www.summarizebot.com/
[3] SMMRY, "Summarize articles, text, websites, essays and documents." [Online]. Available: https://smmry.com/
[4] TLDRThis, "Tldr this." [Online]. Available: https://www.tldrthis.com/
[5] TextCompactor, "Text compactor." [Online]. Available: https://www.textcompactor.com/

Related Work : Literature

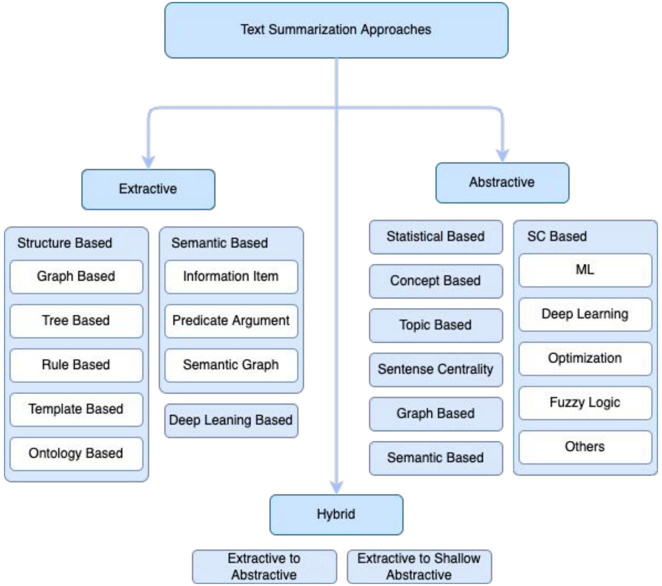
- **El-Kassas et al. [6]** have classified the summarization systems considering different aspects as shown here



[6] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic text summarization: A comprehensive survey,” Expert Systems with Applications, vol. 165, p. 113679, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420305030

Related Work : Literature

- Abstractive, Extractive and combination of both named Hybrid Text summarization approaches shown here as mentioned by **El-Kassas et al. [6]**.



Text Summarization Approaches

Extractive

Structure Based
Graph Based
Tree Based
Rule Based
Template Based
Ontology Based

Semantic Based
Information Item
Predicate Argument
Semantic Graph

Deep Leaning Based

Abstractive

Statistical Based
Concept Based
Topic Based
Sentense Centrality
Graph Based
Semantic Based

SC Based
ML
Deep Learning
Optimization
Fuzzy Logic
Others

Hybrid

Extractive to Abstractive

Extractive to Shallow Abstractive

[6] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," Expert Systems with Applications, vol. 165, p. 113679, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420305030

Related Work : Literature

- **Cohan et al [7]** have proposed a discourse-aware attention model that considers the document’s discourse structure during the summarization process. The discourse structure refers to the document’s organization and coherence, such as the presence of sections, headings, and paragraph transitions. Datasets they have provided as **arXiv** and **Pubmed** collected from respective repositories are a vital resource in this domain.

- **Liu and Lapata [8]** presented a summarization approach that employs pretrained language models, with a particular emphasis on the use of models such as **BERT (Bidirectional Encoder Representations from Transformers) [3]**

[7] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, “A discourse-aware attention model for abstractive summarization of long documents,” arXiv preprint arXiv:1804.05685, 2018.
[8] Liu, Y., Lapata, M.: Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345 (2019)

Related Work : Datasets

Dataset	No of Articles	Details
CNN/Daily Mail dataset [9]	More than 300k	On average, the source documents in the training set contain 766 words across 29.74 sentences, while the summaries comprise 53 words and 3.72 sentences
NYT [10]	More than 1.8 million	Articles published by the New York Times from January 1, 1987, to June 19, 2007
XSum [11]	More than 220k	Consists of BBC articles and accompanying single sentence summaries
Newsroom dataset [12]	1.3 million (1.2 million publicly available)	This contains 1.3 million articles and summaries written by authors and editors in the newsrooms of 38 major publications
Bytecup dataset [13]	1.3 million (1.1 million released for training)	The Byte Cup 2018 International Machine Learning Contest released a new dataset, commonly known as the Bytecup dataset
ArXiv Dataset [7]	215K	Avg. doc length=4938 words; Avg. summary length=220 words
PubMed Dataset [7]	133K	Avg. doc length=3016 words; Avg. summary length=203 words

[7] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, “A discourse-aware attention model for abstractive summarization of long documents,” arXiv preprint arXiv:1804.05685, 2018.

[9] “Github - abisee/cnn-dailymail: Code to obtain the cnn / daily mail dataset (non-anonymized) for summarization,” 2021. [Online]. Available: https://github.com/abisee/cnn-dailymail

[10] “The new york times annotated corpus.” [Online]. Available: https://catalog.ldc.upenn.edu/LDC2008T19

[11] Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. arXiv preprint arXiv:1808.08745.

[12] M. Grusky, “Cornell newsroom summarization dataset,” 2021. [Online]. Available: https://lil.nlp.cornell.edu/newsroom/

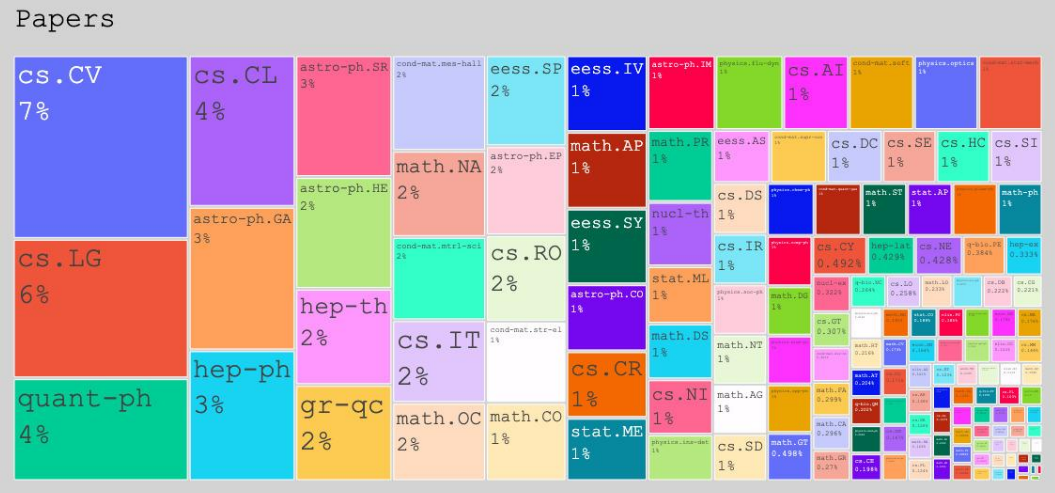
[13] Byte cup 2018 international machine learning contest.” [Online]. Available: https://www.biendata.xyz/competition/bytecup2018/

5. Dataset

Dataset

Evaluating existing datasets like **arXiv** and **Pubmed** we identified that section wise breakdown of research paper text was not found in considerably enough manner.

For this we generated a new dataset from **arXiv** repository with section wise divided text.

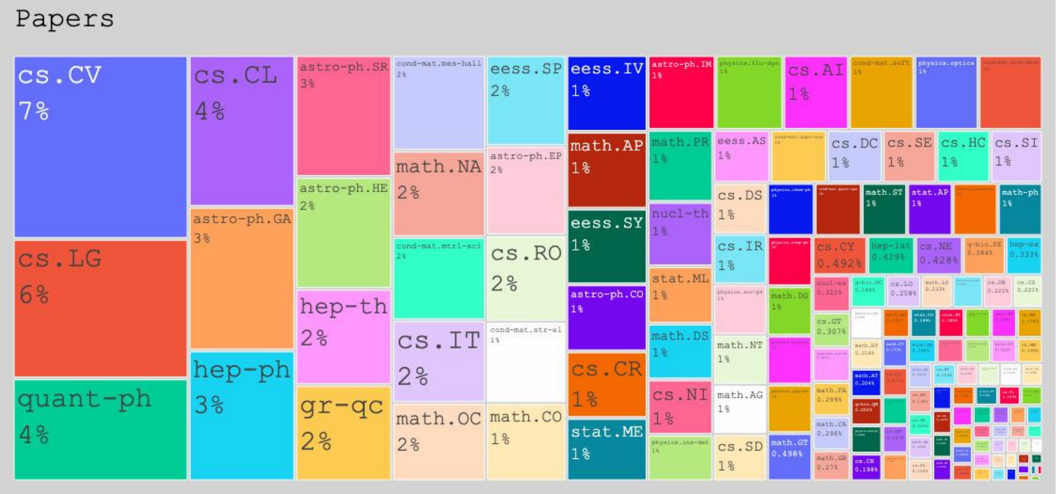


arXiv tags distribution of our dataset

Dataset

For this we generated a new dataset from **arXiv** repository with section wise divided text.

- 300k research papers, section divided
- 350k research papers that were not divided into sections
- Made it publically available

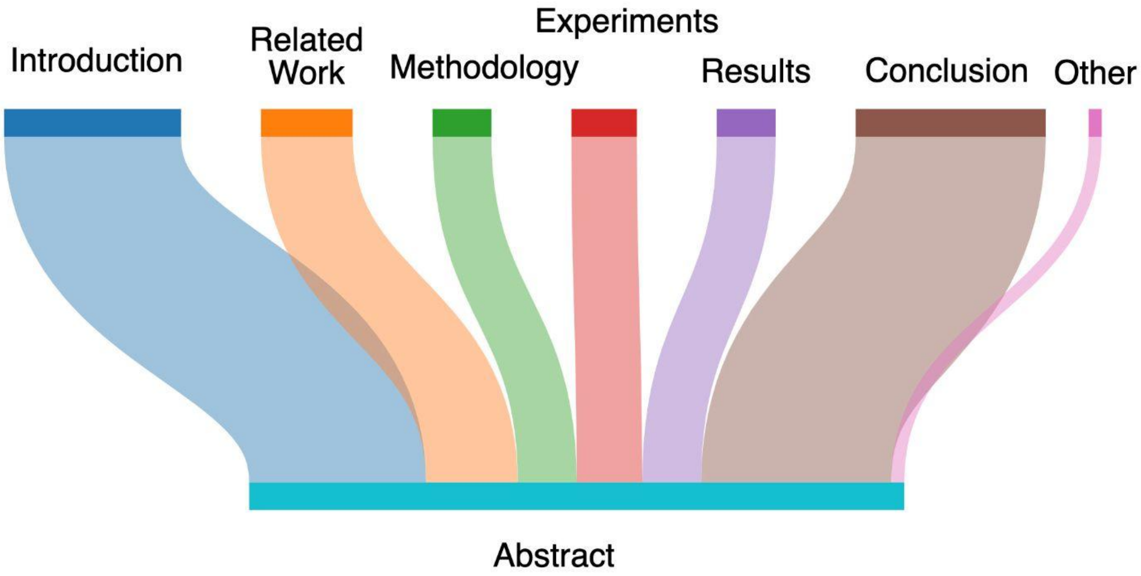


arXiv tags distribution of our dataset

6. Methodology

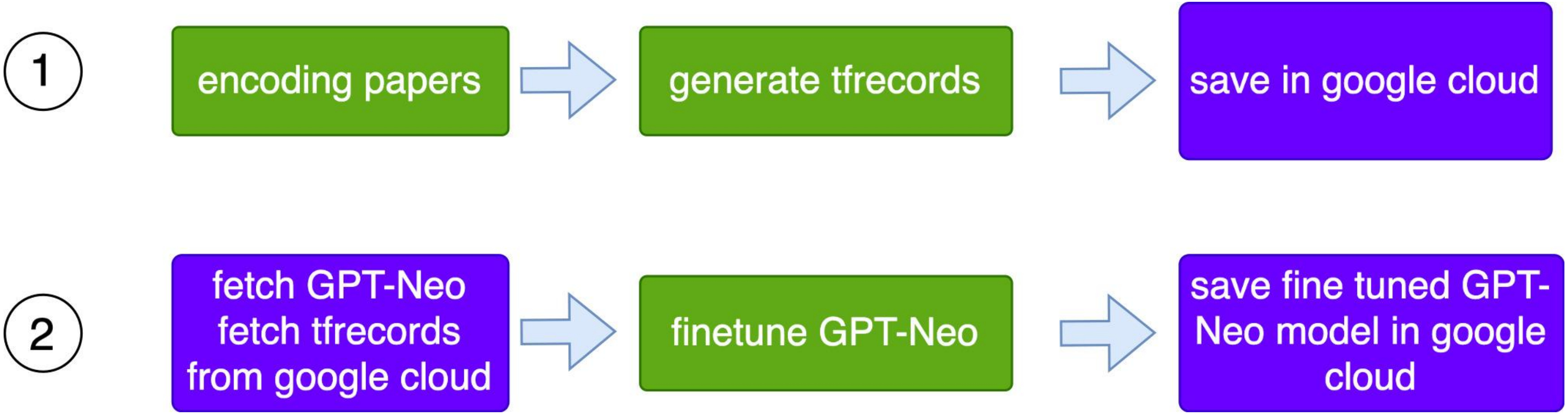
Methodology : Relevance Matrix

Involvement to the abstract section from other sections compared by ROUGE-2 as a score, has given us these values which will be considered as **“Relevance Matrix”**



Introduction	27%
Related Work	14%
Methodology	09%
Experiments	10%
Results	09%
Conclusion	29%
Other	02%

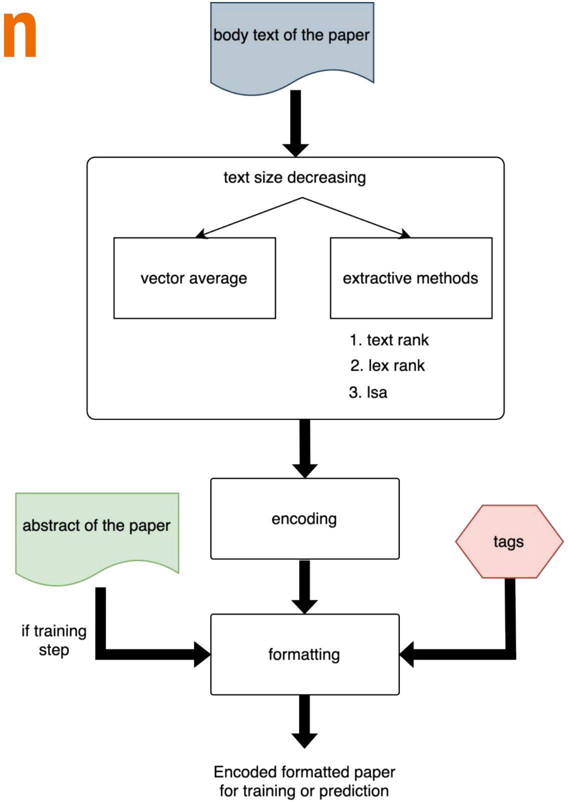
Methodology : Workflow


1
encoding papers
generate tfrecords
save in google cloud
2
fetch GPT-Neo fetch tfrecords from google cloud
finetune GPT-Neo
save fine tuned GPT-Neo model in google cloud

Training flow. **1.** Generation of tfrecords, **2.** Fine Tuning GPT-Neo

Methodology : Presummarization

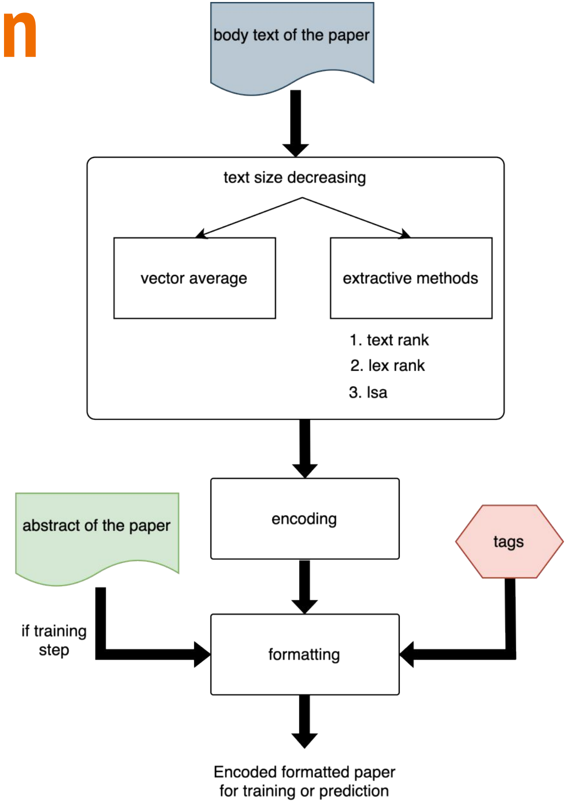
- Evaluated 4 main pre summarization techniques.
  - Vector average
  - Text rank algorithm
  - Lex rank algorithm
  - LSA algorithm



[14] Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).
[15] Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22, 457-479.
[16] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

Methodology : Presummarization

- Evaluated 4 main pre summarization techniques.
  - Vector average
  - Text rank algorithm
  - Lex rank algorithm
  - **LSA algorithm**



[14] Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).
[15] Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22, 457-479.
[16] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

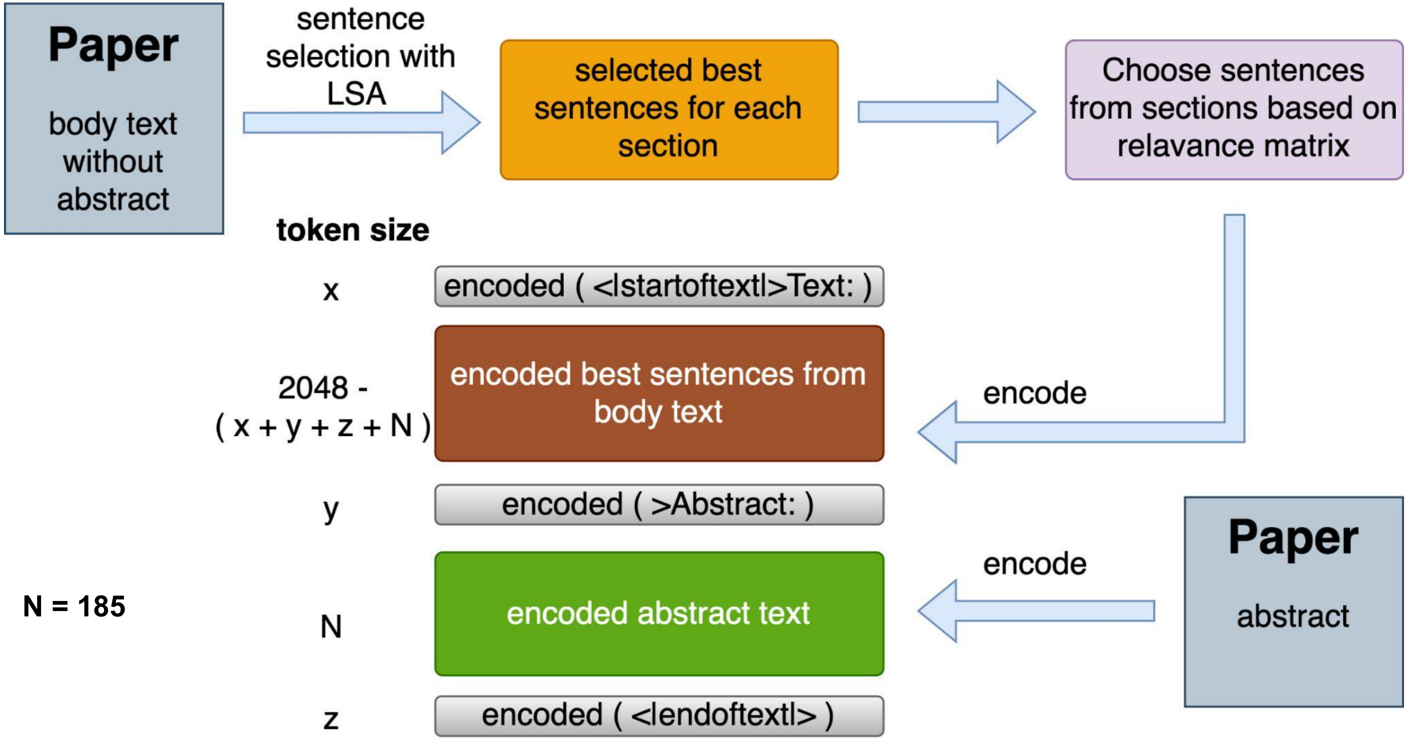
Methodology : LSA - (Latent Semantic Analysis)

How LSA Helps in Summarization

- Step 1: Convert the text into a term-document matrix (TDM).
- Step 2: Apply Singular Value Decomposition (SVD) to identify key topics.
- Step 3: Rank sentences based on their contribution to dominant topics.
- Step 4: Select the most important sentences for the summary.

[16] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

Methodology : Data encoding

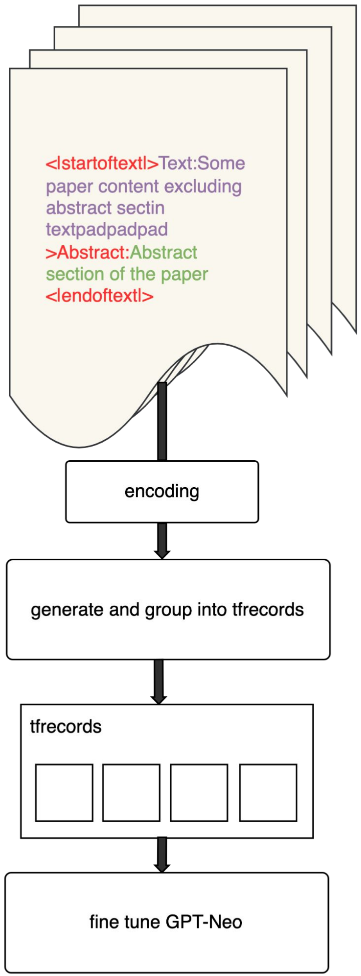


Paper
body text without abstract
sentence selection with LSA
selected best sentences for each section
Choose sentences from sections based on relavance matrix
token size
x
encoded (<|startoftext|>Text:)
2048 - (x + y + z + N)
encoded best sentences from body text
encode
y
encoded (>Abstract:)
N
encoded abstract text
encode
Paper
abstract
z
encoded (<|endoftext|>)

N = 185

Methodology : Training

- After encoding is done data was used to generate tfrecords and fine tuning was done on GPT-Neo



<|startoftext|>Text:Some paper content excluding abstract sectin textpadpadpad >Abstract:Abstract section of the paper <|endoftext|>

encoding

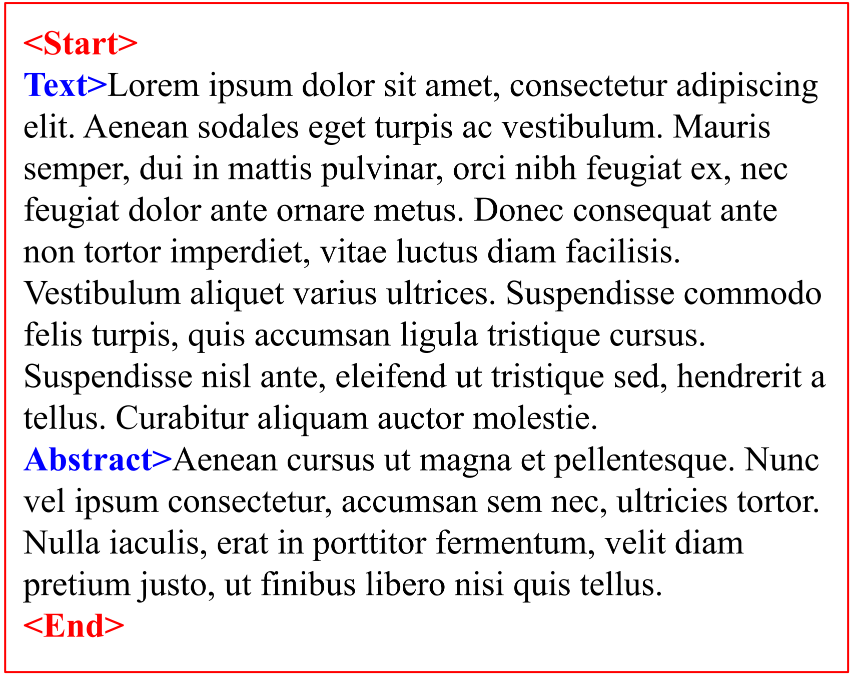
generate and group into tfrecords

tfrecords

fine tune GPT-Neo

Methodology : Training

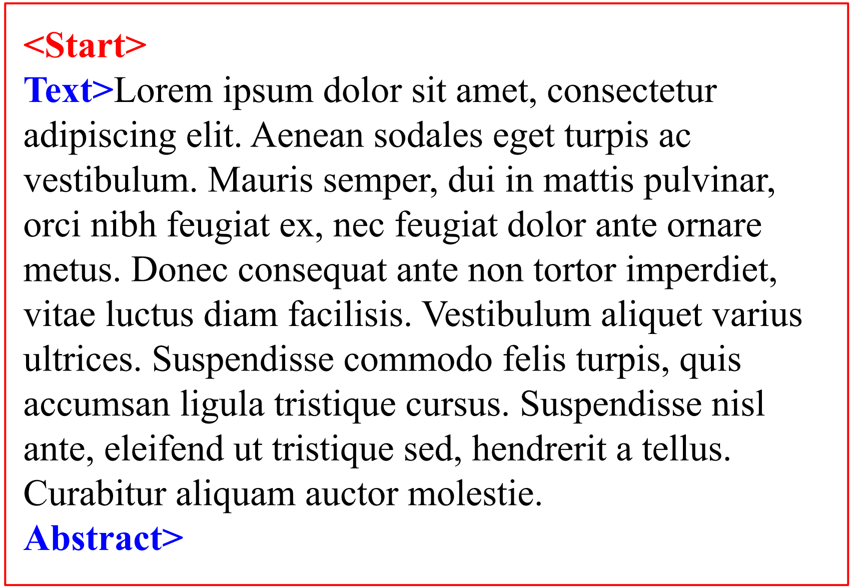
- GPT models are trained with expected outcome from the text
- For prediction we provide the start text without abstract section
- It will predict the rest starting from last word we have given



<Start>
Text>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean sodales eget turpis ac vestibulum. Mauris semper, dui in mattis pulvinar, orci nibh feugiat ex, nec feugiat dolor ante ornare metus. Donec consequat ante non tortor imperdiet, vitae luctus diam facilisis. Vestibulum aliquet varius ultrices. Suspendisse commodo felis turpis, quis accumsan ligula tristique cursus. Suspendisse nisl ante, eleifend ut tristique sed, hendrerit a tellus. Curabitur aliquam auctor molestie.
Abstract>Aenean cursus ut magna et pellentesque. Nunc vel ipsum consectetur, accumsan sem nec, ultricies tortor. Nulla iaculis, erat in porttitor fermentum, velit diam pretium justo, ut finibus libero nisi quis tellus.
<End>

Methodology : Prediction

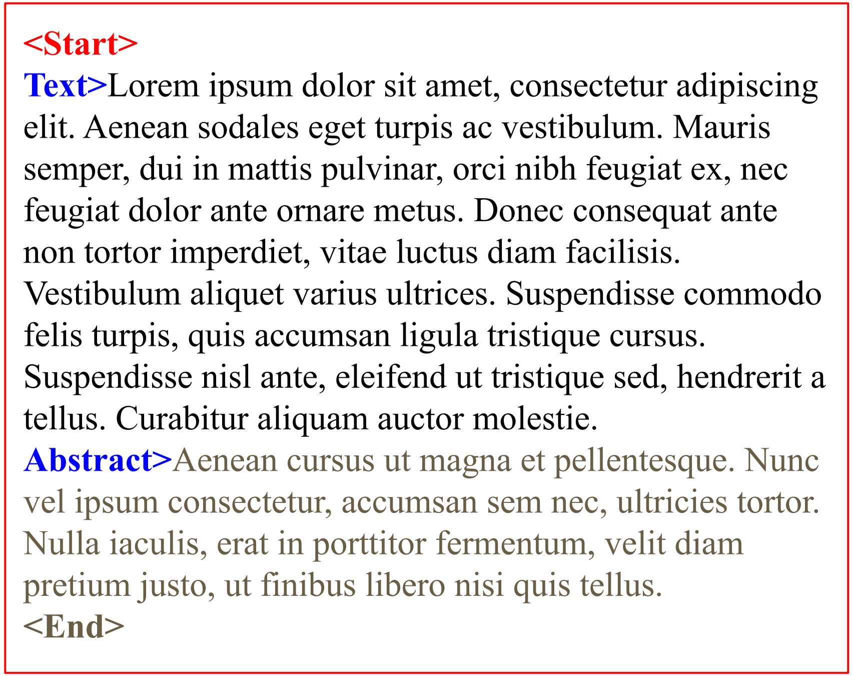
- GPT models are trained with expected outcome from the text
- For prediction we provide the start text without abstract section
- It will predict the rest starting from last word we have given



<Start>
Text>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean sodales eget turpis ac vestibulum. Mauris semper, dui in mattis pulvinar, orci nibh feugiat ex, nec feugiat dolor ante ornare metus. Donec consequat ante non tortor imperdiet, vitae luctus diam facilisis. Vestibulum aliquet varius ultrices. Suspendisse commodo felis turpis, quis accumsan ligula tristique cursus. Suspendisse nisl ante, eleifend ut tristique sed, hendrerit a tellus. Curabitur aliquam auctor molestie.
Abstract>

Methodology : Prediction

- GPT models are trained with expected outcome from the text
- For prediction we provide the start text without abstract section
- It will predict the rest starting from last word we have given

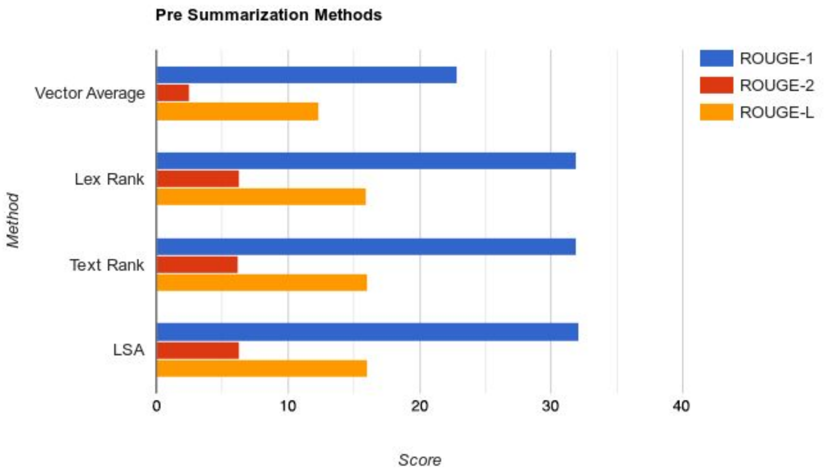


<Start>
Text>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean sodales eget turpis ac vestibulum. Mauris semper, dui in mattis pulvinar, orci nibh feugiat ex, nec feugiat dolor ante ornare metus. Donec consequat ante non tortor imperdiet, vitae luctus diam facilisis. Vestibulum aliquet varius ultrices. Suspendisse commodo felis turpis, quis accumsan ligula tristique cursus. Suspendisse nisl ante, eleifend ut tristique sed, hendrerit a tellus. Curabitur aliquam auctor molestie.
Abstract>Aenean cursus ut magna et pellentesque. Nunc vel ipsum consectetur, accumsan sem nec, ultricies tortor. Nulla iaculis, erat in porttitor fermentum, velit diam pretium justo, ut finibus libero nisi quis tellus.
<End>

7. Results

Results : Pre-Summarization methods

Pre-Summarization Method	R1	R2	RL
Vector Average	22.91	2.50	12.32
Lex Rank	31.93	6.36	15.99
Text Rank	31.96	6.20	16.01
LSA	**32.13**	**6.37**	**16.02**



Results

Dataset	Summarization Method	R1	R2	R3	RL
arXiv	Cohan et al **[58]**	35.80	**11.05**	**3.62**	**31.80**
	This work	33.77	8.73	2.67	18.57
	This work + stemmer	**36.12**	9.56	2.96	19.55
Pubmed	Cohan et al **[58]**	**38.93**	**15.37**	**9.97**	**35.21**
	This work	33.47	8.88	3.39	18.28
	This work + stemmer	35.85	9.64	3.66	19.06
Section divided new dataset	This work + relevance matrix	34.14	7.66	2.37	17.51
	This work + relevance matrix + stemmer	**37.41**	8.51	2.60	18.66

[7] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, “A discourse-aware attention model for abstractive summarization of long documents,” arXiv preprint arXiv:1804.05685, 2018.

8. Conclusion

Conclusion

- Long text summarization faces the challenge of limited token size in LLMs.
- With various ways to deal for this challenge, pre-summarization using extractive models was one of them
- Even with presummaration in a section divided document, the focus for each section is not a single weight.
- For such divided, weighted section pre-summarization we introduce the **relevance matrix** to govern the weighted extractive summarization.
- With this approach we could elevate the performance and the accuracy of summary generation in research papers.

9. Publications

Publications

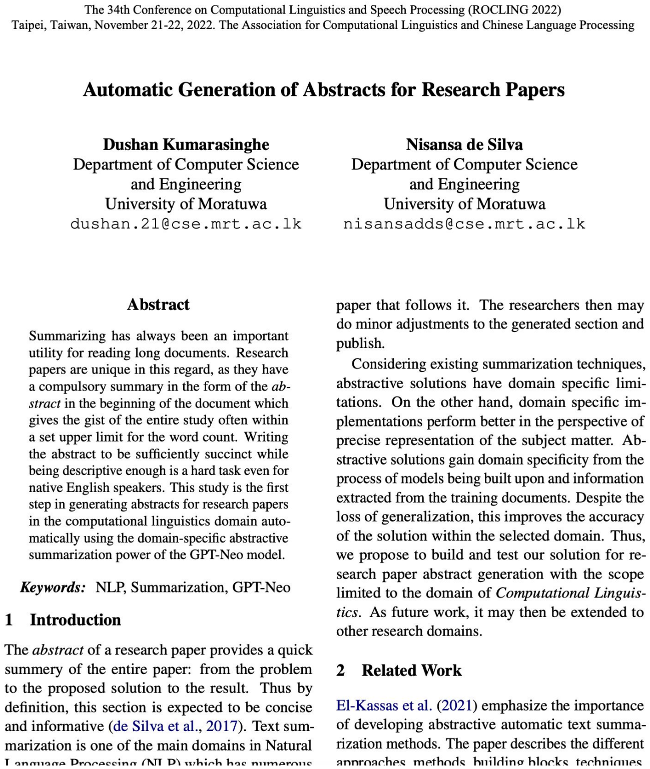
1. Dushan Kumarasinghe, Nisansa de Silva “Automatic Generation of Abstracts for Research Papers” 34th annual Conference on Computational Linguistics and Speech Processing in Taiwan (ROCLING 2022). **[14]**

2. Dushan Kumarasinghe, Nisansa de Silva “Abstract Generation with Hybrid Model Supported by Relevance Matrix” 16th International Conference on Computational Collective Intelligence in Germany (ICCCI 2024). **[15]**

[17] D. Kumarasinghe and N. De Silva, “Automatic generation of abstracts for research papers,” in Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022), 2022, pp. 221–229.
[18] D. Kumarasinghe and N. de Silva, “Abstract generation with hybrid model supported by relevance matrix,” in International Conference on Computational Collective Intelligence. Springer, 2024, pp. 211–223.

Publications : 1 - ROCLING 2022

- Evaluated extractive summarization methods
- Introduced the hybrid method utilizing extractive summarization algorithms with GPT-Neo LLM
- Evaluated the hybrid method performance



The 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)
Taipei, Taiwan, November 21-22, 2022. The Association for Computational Linguistics and Chinese Language Processing

Automatic Generation of Abstracts for Research Papers

Dushan Kumarasinghe
Department of Computer Science
and Engineering
University of Moratuwa
dushan.21@cse.mrt.ac.lk

Nisansa de Silva
Department of Computer Science
and Engineering
University of Moratuwa
nisansadds@cse.mrt.ac.lk

Abstract

Summarizing has always been an important utility for reading long documents. Research papers are unique in this regard, as they have a compulsory summary in the form of the *abstract* in the beginning of the document which gives the gist of the entire study often within a set upper limit for the word count. Writing the abstract to be sufficiently succinct while being descriptive enough is a hard task even for native English speakers. This study is the first step in generating abstracts for research papers in the computational linguistics domain automatically using the domain-specific abstractive summarization power of the GPT-Neo model.

Keywords: NLP, Summarization, GPT-Neo

1 Introduction

The *abstract* of a research paper provides a quick summery of the entire paper: from the problem to the proposed solution to the result. Thus by definition, this section is expected to be concise and informative (de Silva et al., 2017). Text summarization is one of the main domains in Natural Language Processing (NLP) which has numerous

paper that follows it. The researchers then may do minor adjustments to the generated section and publish.

Considering existing summarization techniques, abstractive solutions have domain specific limitations. On the other hand, domain specific implementations perform better in the perspective of precise representation of the subject matter. Abstractive solutions gain domain specificity from the process of models being built upon and information extracted from the training documents. Despite the loss of generalization, this improves the accuracy of the solution within the selected domain. Thus, we propose to build and test our solution for research paper abstract generation with the scope limited to the domain of *Computational Linguistics*. As future work, it may then be extended to other research domains.

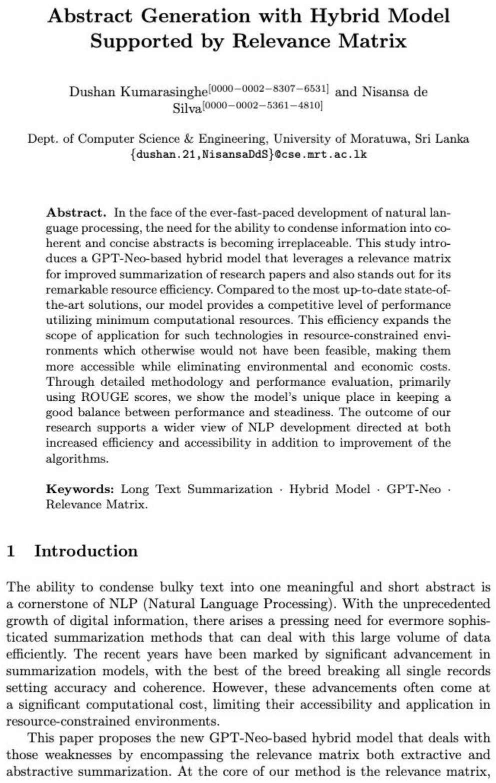
2 Related Work

El-Kassas et al. (2021) emphasize the importance of developing abstractive automatic text summarization methods. The paper describes the different approaches, methods, building blocks, techniques

[17] D. Kumarasinghe and N. De Silva, "Automatic generation of abstracts for research papers," in Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022), 2022, pp. 221–229.

Publications : 2 - ICCCI 2024

- Evaluated abstract formation comparing other sections of the research papers
- Introduced Relevance matrix to tuneup abstract formation with hybrid model utilized extractive algorithms and GPT-Neo model



Abstract Generation with Hybrid Model Supported by Relevance Matrix

Dushan Kumarasinghe$^{[0000-0002-8307-6531]}$ and Nisansa de Silva$^{[0000-0002-5361-4810]}$

Dept. of Computer Science & Engineering, University of Moratuwa, Sri Lanka
{dushan.21,NisansaDdS}@cse.mrt.ac.lk

Abstract. In the face of the ever-fast-paced development of natural language processing, the need for the ability to condense information into coherent and concise abstracts is becoming irreplaceable. This study introduces a GPT-Neo-based hybrid model that leverages a relevance matrix for improved summarization of research papers and also stands out for its remarkable resource efficiency. Compared to the most up-to-date state-of-the-art solutions, our model provides a competitive level of performance utilizing minimum computational resources. This efficiency expands the scope of application for such technologies in resource-constrained environments which otherwise would not have been feasible, making them more accessible while eliminating environmental and economic costs. Through detailed methodology and performance evaluation, primarily using ROUGE scores, we show the model's unique place in keeping a good balance between performance and steadiness. The outcome of our research supports a wider view of NLP development directed at both increased efficiency and accessibility in addition to improvement of the algorithms.

Keywords: Long Text Summarization · Hybrid Model · GPT-Neo · Relevance Matrix.

1 Introduction

The ability to condense bulky text into one meaningful and short abstract is a cornerstone of NLP (Natural Language Processing). With the unprecedented growth of digital information, there arises a pressing need for evermore sophisticated summarization methods that can deal with this large volume of data efficiently. The recent years have been marked by significant advancement in summarization models, with the best of the breed breaking all single records setting accuracy and coherence. However, these advancements often come at a significant computational cost, limiting their accessibility and application in resource-constrained environments.

This paper proposes the new GPT-Neo-based hybrid model that deals with those weaknesses by encompassing the relevance matrix both extractive and abstractive summarization. At the core of our method is the relevance matrix,

[18] D. Kumarasinghe and N. de Silva, “Abstract generation with hybrid model supported by relevance matrix,” in International Conference on Computational Collective Intelligence. Springer, 2024, pp. 211–223.

10. References

References

[1] Resoomer, “Summarizer to make an automatic text summary online.” [Online]. Available: https://resoomer.com/en

[2] SummarizeBot, “Get to know more by reading less!” [Online]. Available: https://www.summarizebot.com/

[3] SMMRY, “Summarize articles, text, websites, essays and documents.” [Online]. Available: https://smmry.com/

[4] TLDRThis, “Tldr this.” [Online]. Available: https://www.tldrthis.com/

[5] TextCompactor, “Text compactor.” [Online]. Available: https://www.textcompactor.com/

[6] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic text summarization: A comprehensive survey,” Expert Systems with Applications, vol. 165, p. 113679, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420305030

[7] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, “A discourse-aware attention model for abstractive summarization of long documents,” arXiv preprint arXiv:1804.05685, 2018.

[8] Liu, Y., Lapata, M.: Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345 (2019)

References

[9] “Github - abisee/cnn-dailymail: Code to obtain the cnn / daily mail dataset (non-anonymized) for summarization,” 2021. [Online]. Available: https://github.com/abisee/cnn-dailymail

[10] “The new york times annotated corpus.” [Online]. Available: https://catalog.ldc.upenn.edu/LDC2008T19

[11] Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. arXiv preprint arXiv:1808.08745.

[12] M. Grusky, “Cornell newsroom summarization dataset,” 2021. [Online]. Available: https://lil.nlp.cornell.edu/newsroom/

[13] Byte cup 2018 international machine learning contest.” [Online]. Available: https://www.biendata.xyz/competition/bytecup2018/

[14] Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).

[15] Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22, 457-479.

[16] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

References

[17] D. Kumarasinghe and N. De Silva, “Automatic generation of abstracts for research papers,” in Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022), 2022, pp. 221–229.

[18] D. Kumarasinghe and N. de Silva, “Abstract generation with hybrid model supported by relevance matrix,” in International Conference on Computational Collective Intelligence. Springer, 2024, pp. 211–223.

Thank You.

—

Q & A