University of Wisconsin – Milwaukee

Department of Computer Science

COMPSCI 723 (Natural Language Processing)

Grad Term Project Report

Project Title:

Implementing Extractive and Abstractive Text Summarization Techniques to Enhance Student Comprehension

Group Members:

Shivam Jayeshkumar Mehta (sjmehta@uwm.edu)

Atharva Pradeep Vaishnav (vaishna2@uwm.edu)

Venkata Kailash Tanniru (vtanniru@uwm.edu)

Semester:

Fall 2024

Instructor:

Dr. Susan Mcroy

1. Introduction

In the current academic landscape, college students are often overwhelmed with vast amounts of textual information from textbooks, research papers, lecture notes, and supplementary materials. This information overload can make it challenging to identify key concepts, retain critical information, and prepare effectively for exams and assignments. The cognitive effort required to process lengthy academic texts often leads to frustration, reduced comprehension, and inefficient learning (Thiede & Anderson, 2003; Chen & Yang, 2022).

Our project, "Implementing Extractive and Abstractive Text Summarization Techniques to Enhance Student Comprehension," aims to address this challenge by developing automated tools that can generate concise and informative summaries of educational content. By implementing both extractive summarization using the TextRank algorithm (Mihalcea & Tarau, 2004) and abstractive summarization using the T5 Transformer model (Raffel et al., 2020), we aim to provide students with summaries that not only reduce the reading load but also enhance comprehension and retention. These tools can help students quickly grasp essential ideas, making learning more efficient and less overwhelming.

2. Real-World Significance

In an era where information overload is a significant challenge for college students, efficient tools for condensing and understanding educational content are increasingly necessary. Studies have shown that when students are faced with too much information, their ability to extract and retain key concepts diminishes, leading to frustration and poor learning outcomes (Chen & Yang, 2022). This is particularly problematic in higher education, where students are expected to digest complex scientific articles, technical documents, and detailed lecture notes within limited timeframes.

Text summarization provides a practical solution to this issue by condensing large volumes of text into shorter, more digestible summaries. Extractive summarization, such as the TextRank algorithm, preserves the original wording and factual accuracy of the content, making it suitable for technical subjects where precision is critical (Nenkova & McKeown, 2012). On the other hand, abstractive summarization models like T5 generate summaries by rephrasing content, resulting in more fluent and coherent text that is often easier to understand (See et al., 2017). This dual approach allows us to cater to different learning needs: students who need precise information retrieval and those who benefit from simplified, paraphrased content.

The significance of this project lies in its potential to improve student comprehension, retention, and learning efficiency. By reducing the cognitive load associated with processing lengthy texts, summarization tools can help students focus on understanding core concepts rather than sifting through extraneous details. For example, a well-designed summarization tool can aid students in preparing for exams by highlighting the most important information, thereby facilitating more effective study sessions (Thiede & Anderson, 2003). Furthermore, readability metrics like the Flesch-Kincaid Grade Level ensure that the summaries are tailored to the appropriate reading level, making complex subjects more accessible (Collins-Thompson, 2014).

The integration of summarization tools into educational platforms, research tools, and content aggregation websites can significantly enhance the learning experience for college students. Whether it is simplifying a dense physics paper or summarizing a lengthy lecture on biology, these tools provide a way to make academic content more manageable and engaging. By leveraging the strengths of both extractive and abstractive summarization, our project offers a versatile solution that addresses the diverse needs of modern learners.

This work builds on established research in NLP and text summarization (Nenkova & McKeown, 2012; Mihalcea & Tarau, 2004; Raffel et al., 2020), contributing to the ongoing effort to improve educational tools through automated summarization technologies.

3. Tools, Technologies, and Data

To implement and compare extractive and abstractive summarization techniques, we utilized a range of tools and libraries commonly used in Natural Language Processing (NLP). These tools enabled us to preprocess data, develop summarization models, and evaluate the quality of the generated summaries.

1. Programming Language:

Python: Selected for its versatility and the rich ecosystem of NLP libraries and frameworks.

2. Libraries and Frameworks

NLTK (Natural Language Toolkit): Used for tokenization, stopword removal, and sentence segmentation.
SpaCy: Assisted in preprocessing tasks like tokenization and lemmatization where needed.
Regular Expressions (re module): For text cleaning tasks, including removing LaTeX formatting, URLs, and citations.
Scikit-Learn: Utilized for TF-IDF vectorization to represent sentences numerically.
NetworkX: Implemented the TextRank algorithm by constructing sentence similarity graphs and applying the PageRank algorithm.
Cosine Similarity: Calculated similarity scores between sentences to build the graph.
Hugging Face Transformers: Implemented the T5 (Text-to-Text Transfer Transformer) model for generating abstractive summaries.
PyTorch: Leveraged for model execution and GPU support, enabling efficient summarization of long texts.
SentencePiece: Tokenization library integrated with the T5 model.
Rouge-score: This is used to calculate ROUGE-1 and ROUGE-2 scores to measure n-gram overlap between generated and reference summaries.
bert_score: Utilized to assess the semantic similarity between generated and reference summaries.
Textstat: Measured the readability of summaries to ensure they align with the comprehension levels of college students by implementing Flesch-Kincaid Grade Level.
Pandas: For efficient data manipulation and processing.
NumPy: For numerical operations and array manipulations.

3. Dataset

For this project, we created a custom JSON-formatted dataset designed to focus on educational content relevant to college students. Our dataset leverages articles from the CCDV Arxiv Summarization Dataset, which is available on Kaggle.
Source: CCDV Arxiv Summarization Dataset

The dataset contains full-text scientific articles and their abstracts across various subjects, including physics, biology, and computer science.

Custom Dataset Creation:

- **Selection of Articles:**

We selected a subset of articles from the CCDV Arxiv Summarization Dataset that are appropriate for college-level education.

- **Learning-Focused Summaries:**

For each selected article, we created custom learning-focused summaries designed to enhance comprehension and retention. These summaries aim to:

i. Highlight key concepts and essential information.
ii. Simplify complex ideas while preserving core meaning.
iii. Align with the reading levels and learning objectives of college students.

- **JSON Format:**

Our dataset is structured in a JSON format, where each entry contains:

i. "article": The full text of the scientific article.
ii. "abstract": Our custom learning-focused summary, which is tailored for educational purposes.

4. Methods

Extractive Summarization Using TextRank

Extractive summarization aims to generate summaries by directly selecting important sentences from the original text. For this project, we implemented the TextRank algorithm, an unsupervised graph-based method inspired by Google's PageRank algorithm (Mihalcea & Tarau, 2004). In TextRank, sentences in a document are represented as nodes in a graph, and edges between nodes reflect the similarity between sentences. The similarity is typically calculated using cosine similarity of the TF-IDF (Term Frequency-Inverse Document Frequency) vector representations of the sentences. Sentences with higher centrality scores, indicating greater importance within the document, are selected to form the summary.

The primary benefit of TextRank is that it preserves the original wording and factual accuracy of the content, making it particularly useful for technical and scientific texts where precise information is critical (Nenkova & McKeown, 2012). Additionally, it is an unsupervised method that does not require labeled data, making it computationally efficient and adaptable to different domains without the need for model training. However, a key limitation of TextRank is that it may produce summaries that lack coherence and fluency since the sentences are extracted verbatim. This can sometimes lead to abrupt transitions and redundancy, especially in lengthy documents. Furthermore, it may struggle with paraphrasing or simplifying complex ideas, which is essential for enhancing comprehension in educational contexts.

The process involves the following steps:

i. Preprocessing: The input text is cleaned and tokenized into sentences. Stopwords and punctuation are removed to focus on meaningful words.
ii. Sentence Representation: Sentences are converted into numerical representations, often using TF-IDF. TF-IDF weights words based on their frequency in the document relative to their frequency in the entire dataset, emphasizing important words while reducing the impact of common words.
iii. Similarity Calculation: The similarity between sentences is calculated, typically using cosine similarity. This measures the angle between the TF-IDF vectors of two sentences, with higher values indicating greater similarity.
iv. Graph Construction: A graph is constructed where each node represents a sentence, and edges represent the similarity scores between sentences.
v. Ranking Sentences: The PageRank algorithm is applied to the graph to rank the sentences based on their importance within the text.
vi. Summary Generation: The top-ranked sentences are selected to form the summary. The number of sentences selected can be controlled by specifying word or sentence limits.

Benefits of Extractive Summarization:

- Preserves factual accuracy by retaining the original wording.
- Domain-agnostic and does not require labeled training data.
- Computationally efficient compared to abstractive methods.

Limitations of Extractive Summarization:

- Can produce summaries with redundancy and lack of coherence.

- Does not paraphrase or simplify complex ideas.
- Less effective for texts requiring contextual understanding.

Abstractive Summarization Using T5 Transformer

Abstractive summarization generates summaries by rephrasing and synthesizing information from the original text. For this project, we implemented the T5 (Text-to-Text Transfer Transformer) model, a state-of-the-art Transformer-based architecture developed by Google (Raffel et al., 2020). The T5 model converts all NLP tasks, including summarization, into a text-to-text format. It uses an encoder-decoder architecture with self-attention mechanisms to understand the input context and generate coherent and fluent summaries. The input text is tokenized, passed through the encoder, and then decoded to produce a summary that captures the essence of the original content.

The main advantage of the T5 model is its ability to produce human-like, coherent, and fluent summaries by paraphrasing the content. This makes it particularly effective for simplifying complex information and improving readability (See et al., 2017). The model is also highly adaptable and can be fine-tuned for specific domains or tasks, enhancing its performance on educational materials. However, abstractive summarization is computationally intensive and requires significant resources, especially for long texts. Additionally, it may occasionally generate hallucinated information—content that is not present in the original text—which can reduce factual accuracy (Dong et al., 2019). Despite these limitations, abstractive summarization is valuable for contexts where readability and comprehension are paramount, such as educational tools designed for college students.

The process involves the following steps:

i. Preprocessing: Clean the input text and tokenize it into subwords using a tokenizer
ii. Encoding: The input text is converted into numerical representations (embeddings) by the model's encoder. The encoder captures the context and meaning of the input sequence.
iii. Decoding: The decoder generates the summary by predicting the following word in the sequence based on the encoded context and previously generated words. The output is iteratively generated until the end-of-sequence token is produced.
iv. Beam Search: To improve the quality of the generated summaries, beam search is often used. This technique keeps track of multiple possible sequences at each step and selects the most likely one.

Benefits of Abstractive Summarization:

- Produces fluent, coherent, and human-like summaries.
- Can simplify complex content, making it easier to understand.
- Capable of paraphrasing and generalizing information.

Limitations of Abstractive Summarization:

- Computationally intensive and requires significant resources.
- May generate hallucinated information that is not present in the original text.
- Requires large amounts of data for training and fine-tuning.

Evaluation Metrics

To assess the quality of the summaries generated by both methods, we used the following evaluation metrics: ROUGE, BERTScore, and Flesch-Kincaid Grade Level. These metrics provide a comprehensive evaluation of factual accuracy, semantic similarity, and readability.

1. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE is a widely used metric for automatic summarization evaluation (Lin, 2004). Specifically, we employed ROUGE-1 and ROUGE-2, which measure the overlap of unigrams (single words) and bigrams (two-word phrases), respectively, between the generated summary and the reference summary. ROUGE provides a quantitative measure of how much critical information from the original text is captured in the summary. It is particularly useful for evaluating extractive summaries, which tend to retain the original wording. However, ROUGE has limitations in capturing paraphrased content, making it less effective for evaluating abstractive summaries.

2. BERTScore

BERTScore evaluates the semantic similarity between the generated summary and the reference summary using contextual embeddings from the BERT (Bidirectional Encoder Representations from Transformers) model (Zhang et al., 2019). Unlike ROUGE, which relies on exact word overlap, BERTScore can capture nuanced meaning and paraphrased content, making it ideal for evaluating abstractive summaries. BERTScore computes precision, recall, and F1 scores based on the similarity of word embeddings, providing a more robust measure of content quality. While BERTScore is powerful, it is computationally intensive and may require significant resources for large datasets.

3. Flesch-Kincaid Grade Level

The Flesch-Kincaid Grade Level measures the readability of a text by estimating the U.S. school grade level required to understand it (Kincaid et al., 1975). This metric is particularly relevant for educational content, as it helps ensure that the summaries are appropriate for the target audience—college students. Summaries with lower Flesch-Kincaid scores are easier to read, while higher scores indicate more complex text. This metric is beneficial for comparing the readability of extractive and abstractive summaries. However, it does not account for content accuracy or coherence, making it necessary to use alongside other metrics like ROUGE and BERTScore.

5. Evaluations and Results

Quantitative Evaluation Results:
The table below presents the quantitative evaluation results using ROUGE, BERTScore, and Flesch-Kincaid Grade Level for both summarization methods:

	Method	Avg_ROUGE-1	Avg_ROUGE-2	Avg_BERTScore_F1	Avg_Flesch_Kincaid
0	Extractive	0.326505	0.101785	0.832162	19.77
1	Abstractive	0.267402	0.045906	0.832388	10.29

Interpretation of Quantitative Results:

- ROUGE Scores:
  - Extractive Summarization has higher ROUGE-1 and ROUGE-2 scores compared to Abstractive Summarization. This is expected as extractive methods directly select sentences from the original text, leading to greater word overlap.
  - The lower ROUGE scores for abstractive summarization indicate that the generated summaries use paraphrasing, reducing exact word matches with the reference abstract.
- BERTScore:
  - Both methods achieve similar BERTScore F1 values (~ 0.83), suggesting that both types of summaries maintain the core semantic meaning of the original text.
- Flesch-Kincaid Grade Level:
  - Extractive Summaries have a higher Flesch-Kincaid Grade Level (19.77), making them suitable for advanced readers.
  - Abstractive Summaries are easier to read with a Flesch-Kincaid Grade Level of 10.29, making them more accessible for college students.

Qualitative Evaluation Results:
The table below presents the qualitative evaluation results based on the scores provided by 10 college students who rated the summaries on a scale of 1 to 10 for clarity, relevance, and engagement

Name	Extractive Score (Out of 10)	Abstractive Score (Out of 10)
Nachiket Paradkar	6	9
Soham Shah	6	6
Vishvak Bhatt	8	6
Jenish Patel	7	6
Nikunj Bhat	7	7
Ankit Parikh	7	7
Kavish Shah	6	10
Darpan Patel	10	10
Devansh Shah	6	6
Parth Kapadiya	10	10
Average	7.3	7.7

Interpretation of Qualitative Results

- Comprehensibility and Engagement: Abstractive Summaries scored slightly higher on average (7.7) compared to Extractive Summaries (7.3), indicating that students found abstractive summaries to be more readable and engaging.
- Preference for Simplicity: Students preferred abstractive summaries when they needed simplified explanations and better readability.
- Preference for Accuracy: Extractive summaries were preferred when the original wording and factual accuracy were crucial, especially for technical details.

Generated Summary Sample Output:

- Reference Summary:

"A new barrier penetration formula, derived from the WKB approximation, accounts for the long-range Coulomb interaction, making it effective for low-energy penetrations below the Coulomb barrier. The formula provides accurate penetrability calculations and closely matches WKB results. Applied to alpha-decay half-life evaluations for atomic nuclei, it demonstrates excellent agreement with experimental data, particularly for spherical nuclei. This formula enhances the study of barrier penetrability in nuclear processes and fusion reactions."

- Extractive Summary:

"in the present work , we derived a new barrier penetration based on the wkb approximation . in the study of barrier penetration in nuclear physics , the parabolic approximation is usually adopted because an analytical solution exists for the penetrability of a parabola barrier potential . in the present work , we derived a new barrier penetration , eq . therefore , the new could be used in the study of barrier penetration at energies much smaller than the barrier height ."

- Abstractive summary:

"quantum tunneling is a common quantum phenomenon . it governs many other nuclear processes such as fission and fusion . the tunneling effect is also widely adopted today in the study of heavy ion reactions . the parabolic approximation works remarkably well both for the penetrability and for the fusion cross section ."

Outcomes

From the results, we can draw the following insights:

Extractive Summarization:

- Strengths: Preserves the original wording and factual details, making it suitable for technical subjects where precision is required.
- Weaknesses: Summaries tend to be more complex and harder to read, as reflected by the higher Flesch-Kincaid Grade Level.

Abstractive Summarization:

- Strengths: Produces more readable and engaging summaries by paraphrasing and simplifying content, making it suitable for educational contexts.
- Weaknesses: Occasional loss of precise details and lower ROUGE scores due to paraphrasing.

Significance of Outcomes

The outcomes of our project demonstrate the practical value of integrating both extractive and abstractive summarization techniques to enhance student comprehension. The extractive method, with its high ROUGE scores and preservation of factual accuracy, is ideal for scenarios where retaining precise technical details is essential, such as in scientific and research-oriented materials. On the other hand, the abstractive method, with its lower Flesch-Kincaid Grade Level and higher readability, simplifies complex information and improves accessibility, making it suitable for educational contexts where comprehension and engagement are paramount. The combination of quantitative metrics like ROUGE, BERTScore, and readability scores, along with qualitative feedback from college students, highlights the strengths and limitations of each method. By leveraging both techniques, educational tools can be developed to help students efficiently process large volumes of information, reduce cognitive load, and improve learning outcomes, ultimately making academic resources more accessible and practical.

6. Discussion and Significance

Discussion of Findings

Our project explored the implementation of both extractive and abstractive summarization techniques to assist college students in digesting complex academic materials. The findings from both quantitative and qualitative evaluations provide valuable insights into the strengths and weaknesses of each approach.

Extractive Summarization consistently achieved higher ROUGE-1 and ROUGE-2 scores, reflecting its ability to retain the exact wording and factual accuracy of the original text. This makes it suitable for summarizing technical documents where precision is critical, such as scientific articles and research papers. However, the higher Flesch-Kincaid Grade Level indicates that these summaries are more complex and may not be easily digestible for students who struggle with technical jargon.

Abstractive Summarization produced summaries that were more readable and engaging, as indicated by the lower Flesch-Kincaid Grade Level and higher scores in the qualitative evaluation. The ability of the T5 model to paraphrase and simplify content makes abstractive summarization ideal for educational contexts where comprehension is more important than verbatim accuracy. However, lower ROUGE scores and occasional hallucinations (introducing information not present in the original text) highlight a trade-off between readability and factual precision.

The BERTScore results show that both methods maintain the overall semantic integrity of the original content, confirming that the core message is preserved even when the wording differs.

Significance of the Results

The significance of these results lies in their potential to address the challenges faced by students dealing with information overload. By automating the summarization of academic texts, this project can:

- Enhance Learning Efficiency: Students can quickly grasp key concepts without reading entire papers or chapters, allowing them to allocate more time to critical thinking and analysis.
- Reduce Cognitive Load: Simplified abstractive summaries make complex information more accessible, reducing frustration and improving comprehension for students who struggle with dense academic texts.
- Support Diverse Learning Needs: The dual approach of extractive and abstractive summarization caters to different learning preferences. Students who need technical accuracy can benefit from extractive summaries, while those who prefer simpler explanations can use abstractive summaries.
- Enable Personalized Learning Tools: Educational platforms can integrate these summarization techniques to offer customized summaries based on a student's learning style, academic level, and subject matter.

Broader Impact

The broader impact of this project extends beyond college students. Potential beneficiaries include:

- Researchers: Quickly distill key information from extensive literature reviews.
- Educators: Provide concise study materials and lecture notes to enhance teaching efficiency.
- Professionals: Summarize technical reports, white papers, and business documents to improve decision-making.

Additionally, the integration of summarization tools into e-learning platforms, research databases, and productivity applications can transform the way knowledge is consumed and disseminated.

Limitations

While the results are promising, the project has some limitations:

- Computational Constraints: The T5 model for abstractive summarization is resource-intensive and may not be feasible for real-time processing on low-end hardware.
- Domain-Specific Performance: The performance of both methods can vary depending on the subject matter. Fine-tuning the models for specific domains like medicine or law could improve accuracy.
- Hallucination in Abstractive Summaries: Abstractive methods may generate content that is not factually present in the original text, which can be problematic in contexts requiring high accuracy.

Future Directions

To address these limitations and build on our findings, future work could focus on:

- Hybrid Summarization Techniques: Combining extractive and abstractive approaches to create summaries that are both accurate and readable.
- Domain-Specific Fine-Tuning: Adapting the summarization models to specific academic disciplines to improve relevance and accuracy.
- Real-Time Summarization: Optimizing the computational efficiency of abstractive models for real-time applications on mobile and web platforms.
- User Feedback Integration: Incorporating feedback from students and educators to continuously improve the quality and effectiveness of the summaries.

By addressing these areas, the project can evolve into a robust tool that enhances the learning experience, reduces information overload, and supports knowledge acquisition across various domains.

7. Conclusion

In this project, we developed and evaluated automated extractive and abstractive summarization techniques to aid college students in comprehending complex academic materials. Using the TextRank algorithm for extractive summarization and the T5 Transformer model for abstractive summarization, we aimed to address the challenge of information overload by providing concise, accurate, and readable summaries.

Our findings demonstrate that extractive summarization excels in maintaining factual accuracy and is particularly effective for technical subjects where precision is crucial. However, the complexity of the language in these summaries may hinder comprehension for students who are not well-versed in the subject matter. On the other hand, abstractive summarization generates more readable and engaging summaries by paraphrasing and simplifying content, making it better suited for general educational purposes. The trade-off is a slight loss in precision and the occasional generation of content not present in the original text.

Through quantitative evaluation using ROUGE, BERTScore, and Flesch-Kincaid Grade Level, as well as qualitative feedback from college students, we confirmed that both methods have unique strengths. Extractive summarization is ideal for tasks demanding accuracy, while abstractive summarization enhances readability and comprehension. These results highlight the potential of a hybrid approach that combines the strengths of both methods to create optimal learning tools.

The significance of this project lies in its ability to improve learning efficiency, reduce cognitive load, and support personalized education. By automating the summarization process, students can more effectively engage with academic content, saving time and effort. The outcomes also pave the way for integrating these techniques into educational platforms, research tools, and productivity applications, benefiting not only students but also researchers, educators, and professionals.

Future work can focus on optimizing these models for real-time use, fine-tuning them for specific domains, and developing hybrid methods to balance accuracy and readability. This project is a step forward in using Natural Language Processing to make academic resources more accessible, efficient, and tailored to the diverse needs of learners.

8. References

1. Chen, Y., & Yang, Z. (2022). Information overload and its impact on learning outcomes: A review. Journal of Educational Technology, 35(2), 123-140.
2. Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM) (pp. 91-100).
3. Dong, Y., Wang, S., & Lapata, M. (2019). Towards content transfer through grounded text generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 2626-2636).
4. Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report 8-75. Naval Technical Training Command.
5. Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004) (pp. 74-81).
6. Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 404-411).
7. Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In C. Sammut & G. I. Webb (Eds.), Encyclopedia of Machine Learning (pp. 1010-1017). Springer.
8. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67.
9. See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 1073-1083).
10. Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(4), 618-624.
11. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. In Proceedings of the International Conference on Learning Representations (ICLR).