Text Summarizer for Research Papers

Sumukhi Ganesan
LeAnn Marie Mendoza
Sindhu Swaroop

Introduction to Text Summarizers

Text summarization is the task of producing a concise and fluent summary of a text document while preserving its key information and overall meaning.

There are two main approaches to text summarization:

- **Extractive summarization** methods **identify important sentences or phrases from the text.**
- **Abstractive summarization** attempts to **learn the context, semantics and relationships of different parts of a text**.

Research paper summarization is a unique task in the realm of summarization due to the dense, technical, and multifaceted nature of scientific literature.

Dataset and Preprocessing

The ScisummNet dataset is a **large annotated corpus of scientific papers and their summaries**, specifically collected for scientific paper summarization tasks.

- Contains over 1000 research papers and their corresponding manual summaries
- derived from the ACL anthology network as part of the CL-SciSumm project by the Yale LILY Lab.

Parse XML files and apply regular cleaning and text filters to remove the title, whitespaces, hyperlinks, punctuations and other unnecessary characters and converted to lowercase for standard processing.

Objective

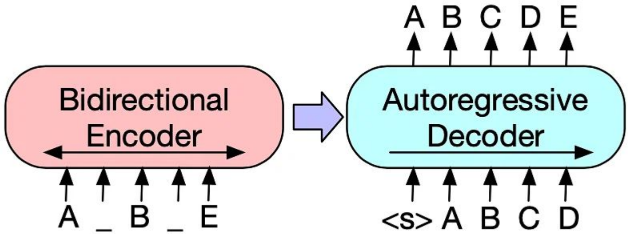
In this project, we will explore these text summarization approaches for research paper summarization:

1. Pre-trained BART (SOTA Transformer)
2. Fine-tuned BART (ScisummNet Data)
3. TextRank (Extractive Approach)
4. Clustering (Extractive Approach)

Transformer based Summarizer

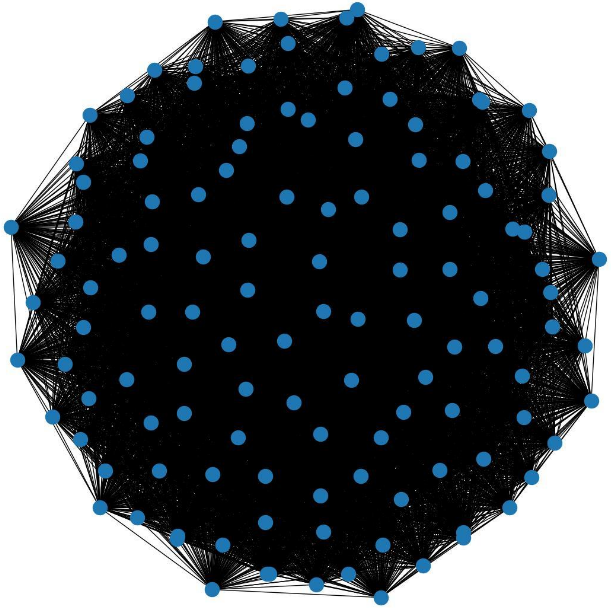
Bidirectional and Auto-Regressive Transformers (BART) developed by Facebook AI and is a SOTA text summarizer.

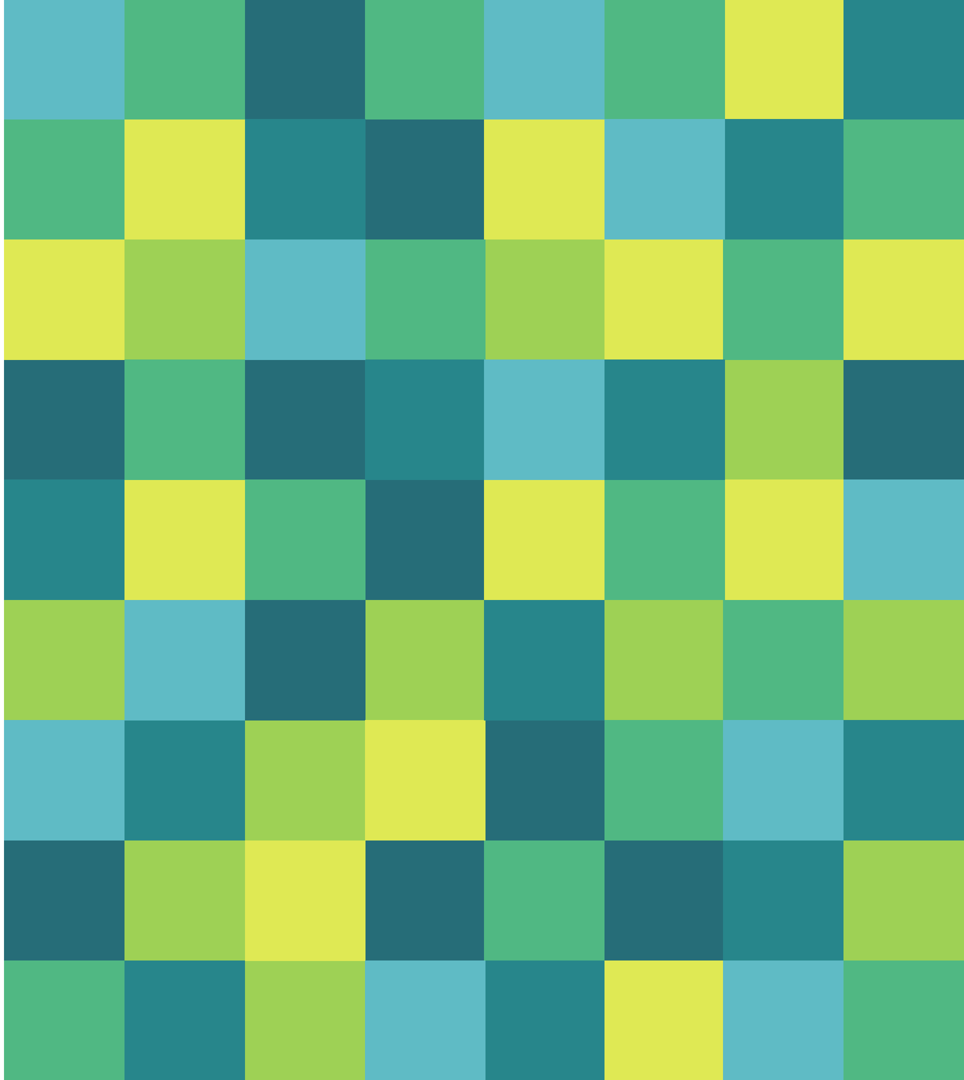
- First input text by **masking out certain words or shuffling the order of sentences**
- Then model attempts to **reconstruct the original from this corrupted version** - allowing BART to learn the context from both tokens.



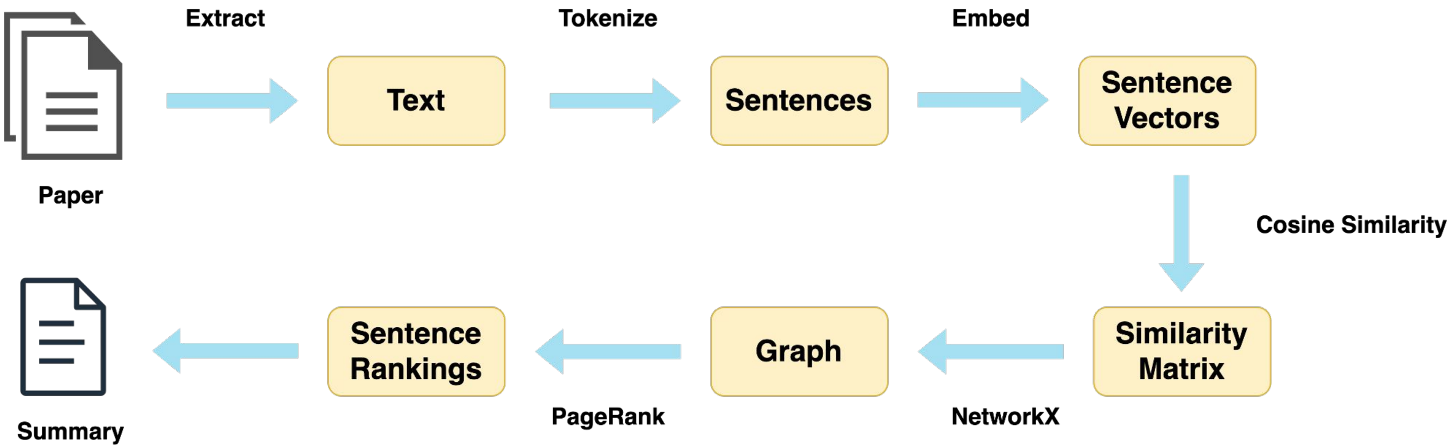
We will utilize both a pre-trained and fine-tuned (trained with ScisummNet) BART model.

Text Rank Summarizer

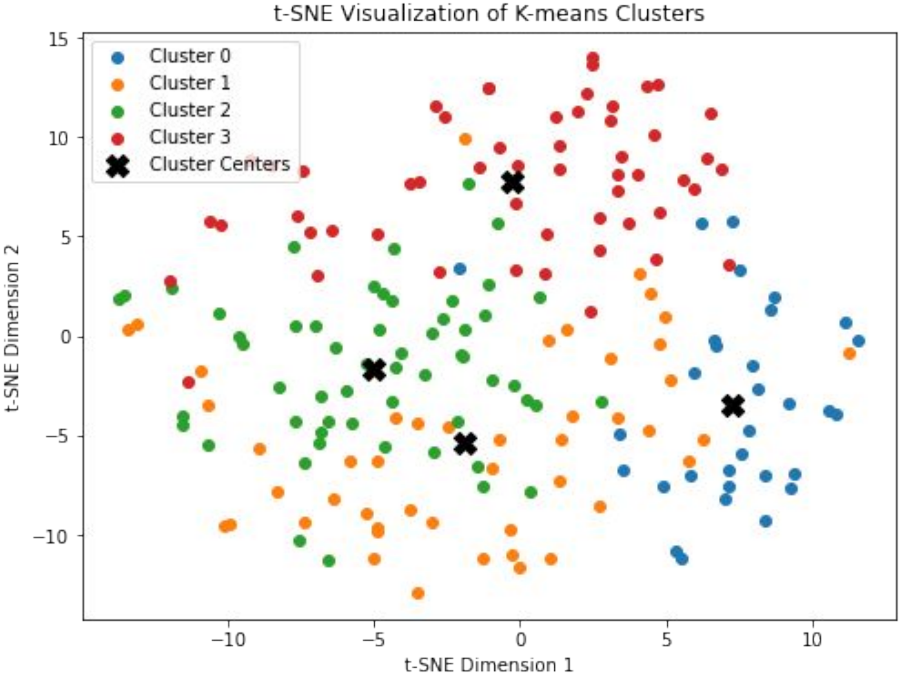




Summarization Flow



Clustering-based Summarizer

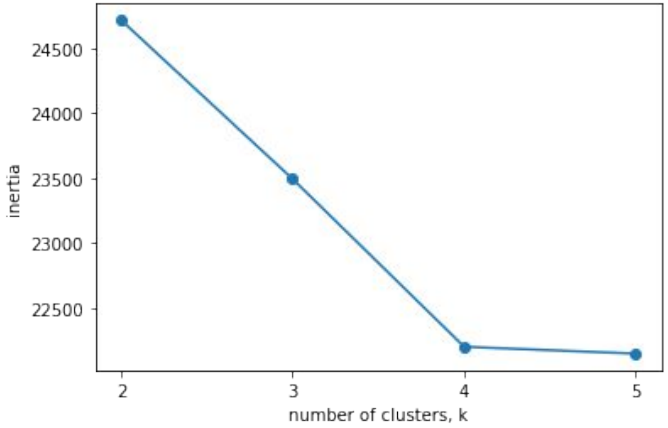


Step 1: Applied **sentence embedding model** "distilbert-base-nli-mean-tokens" to the corpus to extract sentence embeddings

Step 2: Opted for the **K-means clustering** algorithm and employed **elbow method** to determine the ideal number of clusters

Step 3: **Trained** K-means clustering model using our corpus embeddings.

Step 4: Selected a **representative sentence** from each cluster. Collected the representative sentences from each cluster and joined them to form the summary.



Evaluation Metrics and Results

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- **ROUGE-1:** This metric measures the overlap of unigrams (individual words)
- **ROUGE-2:** This metric measures the overlap of bigrams (two consecutive words)
- **ROUGE-L:** This metric measures the overlap of longest common subsequences (LCS).
- **The F1 score** is a measure of both precision and recall.
- 0 is least overlap, 1 is most overlap.

Results

Summarizer	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
Pre-trained BART	0.47	0.36	0.42
Fine-tuned BART	0.50	0.35	0.42
Text Rank	0.34	0.09	0.18
Clustering	0.28	0.11	0.17

Conclusion

Both **pre-trained BART and the ScisummNet fine-tuned BART model performed well** on the ScisummNet dataset, an **expected outcome of utilizing a pretrained model**.

Of the extractive approaches, **TextRank performed the best with the higher unigram ROUGE score**. Beyond the ROUGE scores, the quality of the summaries generated are fairly comparable and have their own advantages and disadvantages.

Future Work

This project currently stands at a **nascent comparison of a few basic summarization algorithms.**

Expansion ideas…

1. more exhaustive survey of text summarization techniques
2. Improve on BART models as they have provided promising accuracy and are easily accessible
3. Explore clustering algorithms such as DBSCAN or OPTICS
4. Investigate density-based or centroid-based approaches for selecting representative sentences from each cluster
5. explore the relevance and possibilities of extractive summarization as it still holds the advantage of being effective with lesser resources.

Thank you