Scalable Multi-Document Text Summarization Using Distributed Deep Learning

1 Introduction

In recent years, the rapid growth of digital content, particularly news articles and research publications, has dramatically increased the demand for efficient summarization tools. At the same time, single-document summarization (SDS) has benefited significantly from advances in deep learning, but multi-document summarization (MDS) - the process of synthesizing content from several related documents into a coherent summary- remains a challenging and less-explored domain. MDS presents unique problems, such as information redundancy, contradictory statements, and varying writing styles across multiple sources, making it inherently complex and computationally demanding.

As more organizations move toward knowledge automation, high-quality summarization models can drastically improve information retrieval efficiency. The rise of large-scale datasets and powerful models makes it timely to investigate scalable solutions for this problem.

This project addresses these challenges by leveraging state-of-the-art transformer-based models and distributed deep learning frameworks to create a scalable MDS system. We aim to demonstrate our methodology and validate it on the Multi-News dataset, a substantial dataset specifically designed for the multi-document summarization problem.

Through distributed training, we seek to make model training feasible and faster for long-sequence inputs. The resulting system could potentially be adapted for downstream tasks such as report generation and document clustering.

2 Problem Formulation

2.1 Description of the Problem

Automatic summarization from multiple documents aims to create concise, informative, and coherent summaries that reflect the content from multiple related texts. Unlike SDS, MDS must handle the redundancy and conflicts arising from multiple perspectives or reporting styles. Achieving accurate, non-redundant, and coherent summaries at scale remains a significant research challenge.

This task also requires models to identify salient information and merge it meaningfully across different source contexts. Furthermore, input formatting and length limitations pose technical hurdles for traditional transformer-based summarizers.

2.2 Big Data Perspective

This project qualifies as a Big Data problem due to the following reasons:

- **Scalability**: The Multi-News dataset, containing over 56,000 instances, requires significant computational resources for processing, training, and evaluation.

- **Complexity of Inputs**: Each summary in Multi-News is generated from multiple documents, resulting in input lengths exceeding typical limits (average input around 2,100 tokens).

- **Computational Intensity**: Training transformer models such as PEGASUS or BART requires substantial GPU resources, ideally addressed through distributed deep learning methodologies.

Our work inherently demands horizontal scaling due to input size, model size, and data volume. The complexity of aligning multiple documents into a shared semantic representation also necessitates distributed processing.

2.3 Impact and Beneficiaries

Developing a robust MDS system has practical implications in various fields:

- **Journalism and Media**: Aggregating news from diverse sources quickly and accurately.

- **Research Communities**: Extracting key findings across multiple academic papers efficiently.

- **Legal and Financial Sectors**: Summarizing extensive reports, legal contracts, and policy documents.

This system can also benefit enterprises that deal with large volumes of internal reports, such as compliance or HR summaries. Educational platforms could use this to generate concise overviews from academic content.

3 Dataset Description

The Multi-News dataset, introduced by Fabbri et al. (2019), will serve as the foundation for this project. It is the first large-scale, human-curated dataset explicitly designed for multi-document summarization. Its well-documented structure and popularity in benchmarking make it ideal for evaluating generalizability and robustness. The dataset aligns well with the project's goal of building scalable, real-world systems.

3.1 Key Statistics

- **Total instances**: 56,216 (training: 44,972; validation: 5,622; test: 5,622)
- **Documents per instance**: Between 2 and 10
- **Average Input Length**: Approximately 2,100 tokens per multi-document cluster
- **Average Summary Length**: Around 260 tokens
- **Diversity of Sources**: Includes content from over 1,500 different news websites, ensuring broad topical and stylistic diversity.
  This heterogeneity introduces useful variation in tone, structure, and redundancy, which makes it suitable for testing generalization and semantic abstraction performance.

3.2 Dataset Features and Challenges

The dataset provides professionally written human summaries, creating a gold-standard reference for evaluating summarization models. Its primary challenge lies in addressing the redundancy and coherence across multiple articles reporting on similar events but from different perspectives. Handling contradicting narratives and reconciling temporal shifts in events presents additional complexity. Summarization models need to effectively balance extraction and abstraction to produce coherent outputs.

4 Methodology and Strategy

4.1 Models and Algorithms

Transformer-based models will be at the project's core, with the following selected for experimentation:

- PEGASUS: Optimized for abstractive summarization.
- BART: Known for robustness with noisy and complex input.
- Longformer Encoder-Decoder (LED): Designed for handling very long sequences efficiently.

These models offer a diverse architectural and pretraining foundation that will help compare performance under different document clustering and merging strategies. We'll analyze trade-offs between compression quality and faithfulness across them.

4.2 Preprocessing and Exploratory Analytics

Data preprocessing and analysis include:

- Cleaning raw textual data (removing HTML tags, normalization).

- Tokenization using model-specific pre-trained tokenizers.
- Exploratory analysis (word frequency distributions, redundancy metrics, summary-input compression ratio).

We will also analyze sentence-level alignment and named entity consistency across merged documents. These insights will inform model input structuring and filtering decisions.

4.3 Distributed Deep Learning Framework

PyTorch with Distributed Data Parallel (DDP) training will facilitate model training across multiple GPUs or nodes. Apache Spark will handle preprocessing and facilitate distributed evaluation and data loading, significantly enhancing scalability.
Distributed inference and evaluation will be implemented using Spark-based pipelines to ensure end-to-end scalability. Checkpointing and logging will be managed using MLflow or equivalent experiment tracking frameworks.

5 Evaluation Methodology and Experimental Design

5.1 Evaluation Metrics

The effectiveness of the summarization will be measured using:

- **ROUGE Scores (1, 2, L)**: Measures textual overlap with reference summaries.
- **BERTScore**: Evaluates semantic similarity and coherence.
- **Extractiveness and Density**: Metrics from the original Multi-News paper, assessing how much of the summary directly extracts from source documents.

5.2 Experimental Setup

Experiments will systematically explore:

- Performance differences between PEGASUS, BART, and LED.
- Impact of different preprocessing and document merging strategies.
- Effects of varying input lengths on summarization quality.

6 Project Timeline

Week	Planned Activities	Responsibility
Week 1	Literature review, Dataset acquisition, Environment setup	Everyone
Week 2	Data preprocessing pipeline implementation and exploratory analysis	Masfiq, Rashidul
Week 3	Distributed training setup, initial PEGASUS model training	Rashidul, Habibullah
Apr 16–21	Model evaluation (ROUGE, BERTScore, Extractiveness)	Rashidul, Habibullah
Week 4	Comparative experiments: PEGASUS, BART, LED	Everyone
Week 4 (Contd.)	Model optimization, results visualization, detailed analysis	Masfiq, Habibullah
Week 5	Report writing, code documentation, project finalization	Everyone

7 Conclusion

This project aims to push the boundaries of current multi-document summarization techniques by leveraging advanced transformer architectures within a scalable, distributed computing environment. By validating methods on the Multi-News dataset, this research seeks to contribute significantly to the fields of automated text summarization and distributed deep learning applications. Our work may lay the foundation for future research on combining retrieval-based and generation-based techniques. The findings could inform design decisions in enterprise document management and intelligent search systems.

Bibliography

1. Fabbri et al., ”Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model,” arXiv:1906.01749, 2019.

2. Zhang et al., ”PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” arXiv:1912.08777, 2019.

3. Lewis et al., ”BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation,” arXiv:1910.13461, 2019.

4. Beltagy et al., ”Longformer: The Long-Document Transformer,” arXiv:2004.05150, 2020.

5. Lin, Chin-Yew, ”ROUGE: A Package for Automatic Evaluation of Summaries,” 2004.