Adaptive Multilingual Document Summarization through Fine-Tuning

Dhananjay Sharma
Stony Brook University
dhananjay.sharma@stonybrook.edu

Parth Pidadi
Stony Brook University
parth.pidadi@stonybrook.edu

Virti Jain
Stony Brook University
virti.jain@stonybrook.edu

Abstract

This document develops a multilingual summarization system using fine-tuned BART and PEGASUS models to generate English summaries from multilingual inputs. It integrates preprocessing, translation, and evaluation into a web application, providing performance comparisons and enhancing accessibility to cross-lingual content.

1 Introduction

Language limitations continue to be a major barrier to international knowledge exchange in an increasingly interconnected world, restricting access to vital information in the public, professional, and academic spheres. Despite the widespread use of large-scale language models, multilingual document summarizing has not advanced much, especially for specialized fields like healthcare, finance, and research. By optimizing cutting-edge multilingual models like mT5, mBART, and Pegasus, this project fills this gap by producing succinct English summaries from papers in multiple languages. This capacity has broad ramifications, facilitating fair access to information and encouraging cross-linguistic cooperation—two things that are critically needed in the age of rapid information sharing.

Its adaptive, domain-specific approach to multilingual summarization, which combines model fine-tuning with performance optimization for practical applications, is what makes this work novel. This project adapts models to domain-specific datasets, assesses their adaptability across linguistic and contextual nuances, and closes the gap between research and usability by providing a workable web-based solution, in contrast to earlier studies that frequently evaluate multilingual summarization generically. By doing this, the project broadens the use of summarization technology to new audiences and goals while simultaneously advancing NLP research and democratizing its applications.

Development and Evaluation of Fine-Tuned Multilingual Summarization Models

Using the MLSUM dataset, this study offers a thorough assessment of sophisticated multilingual models—mT5, mBART, and Pegasus—optimized for document summarizing tasks. The research finds the best configurations for accuracy, efficiency, and domain adaptability by methodically experimenting with model parameters including hidden layers, activation functions, and attention mechanisms. The results give researchers and practitioners useful information about the performance dynamics of multilingual summarization, assisting them in selecting and customizing models for particular use cases.

Creation of a Practical and Comparative Web Application

The project provides an interactive web application that allows users to upload papers in many languages and receive English summaries, as well as to compare model results in order to bridge the gap between research and practical application. Metrics including ROUGE, BLEU, inference time, and computing resource utilization are displayed by the application, which functions as a platform for end users looking for summaries and an instructional tool for NLP practitioners investigating model performance. The increasing demand for interpretable AI systems is addressed by this usability and transparency integration, which makes it pertinent to a wide range of stakeholders.

Promotion of Multilingual NLP Education through Comparative Insights

This project includes a special educational component in recognition of the significance of information transmission in the NLP community. A side panel in the web application illustrates and places the relative advantages and disadvantages of various models. The program helps students and early-career researchers overcome the difficulties of multilingual summarization by breaking down sophisticated NLP metrics into easily understood insights. This contribution guarantees the project's applicability in educational and training contexts by meeting the growing need for real-world, experiential learning materials in data science and natural language processing.

2 Related Work

Rapid progress has been made in multilingual document summarization, especially with the introduction of cross-lingual neural models and extensive datasets. The main goals of current work are to improve the accuracy, scalability, and adaptation of summarization for different languages. Even though these research have made tremendous progress, there are still gaps in the field's ability to customize summarization models for certain domains and maximize their practical utility. This project, in contrast to general cross-lingual frameworks, focuses on optimizing multilingual models like mT5, mBART, and Pegasus for domain-specific datasets and incorporates the final models into a useful, intuitive online application. Our work differs from previous studies due to its real-world application, performance optimization, and adaptability.

2.1 Domain-Specific Fine-Tuning for Multilingual Summarization

Existing projects that offer datasets for training summarization models across several languages include MLSUM (Scialom et al., 2020) and WikiLingua (Ladhak et al., 2020). Although it offers a large corpus for multilingual summarization, MLSUM's application is mostly restricted to general-purpose jobs. Despite its emphasis on abstractive summarization across languages, WikiLingua does not examine domain-specific issues. On the other hand, our project seeks to enhance summarization accuracy and contextual relevance by optimizing cutting-edge multilingual models on domain-specific datasets, including financial reports and medical documents. By expanding multilingual summarization to specialized sectors where linguistic variation frequently poses a challenge, this feature fills a crucial gap.

2.2 Performance Optimization through Comparative Evaluation

Studies like Multilingual Denoising Works such as Attend to the Beginning (Zhong et al., 2020) show improvements in extractive summarization through bidirectional attention mechanisms, while pre-training (Liu et al., 2020) emphasizes the significance of pre-training for better cross-lingual tasks. These studies, however, hardly ever include a thorough comparative analysis of models designed for multilingual summarization. By methodically assessing the performance of mT5, mBART, and Pegasus across factors including hidden layers, activation functions, and attention mechanisms, our work fills this gap. The goal of this thorough comparison is to find the best setups that strike a compromise between computing efficiency and accuracy, offering useful information for future model construction.

2.3 Integration of Research and Practical Application

Although techniques such as Cross-Lingual Abstractive Summarization with Limited Parallel Resources (Dou et al., 2021) successfully handle summarization in situations requiring minimal resources, they are still primarily experimental and do not have real-world applications for end users. By creating an interactive web application that allows people to input documents in multiple languages and receive English summaries, our project extends beyond research. This study also serves as a teaching tool for NLP practitioners by including comparative insights into the application. The project's results are guaranteed to be not just research-focused but also extremely accessible and useful thanks to the integration of usability and transparency.

2.4 Contrast with Existing Works

The majority of current works concentrate on developing models or creating datasets, but they fall short in addressing domain-specific issues or real-world applicability. WikiLingua and MLSUM, for instance, offer basic datasets but do not offer usability testing or domain-specific fine-tuning. Comparative assessments specific to multilingual summarization are also absent from Multilingual Denoising Pre-training and Attend to the Beginning, despite the fact that they present methods for enhancing model performance. With an emphasis on domain adaptation, methodical evaluation, and real-world implementation, our initiative advances the industry by filling these gaps in a unique way.

3 Background

Thanks to developments in large-scale neural models and cross-lingual learning, multilingual summarization has become a crucial field in natural language processing. Extractive summarization, which highlights important sentences or phrases from the original text, was the main emphasis of early attempts. In multilingual environments, methods such as bidirectional attention mechanisms, as examined in Attend to the Beginning (Zhong et al., 2020), showed notable gains in recognizing important information. However, abstractive summarization—where models create summaries by synthesizing information from the input text—has become more popular as a result of extractive techniques' frequent problems with fluency and contextual coherence.

Abstractive summarization has been greatly enhanced by recent multilingual modeling advances like mT5, mBART, and Pegasus. To attain cutting-edge results on cross-lingual tasks, these models make use of transformer topologies and substantial pre-training on multilingual corpora. For example, WikiLingua (Ladhak et al., 2020) aligns multilingual articles for abstractive tasks, while the MLSUM dataset (Scialom et al., 2020) offers a varied corpus for assessing multilingual summarization. Furthermore, pre-training techniques have been shown to be successful in lowering cross-lingual noise, allowing models to function well even with constrained parallel resources. One such technique is Multilingual Denoising Pre-training (Liu et al., 2020).

Even with these developments, current methods are still limited when used for domain-specific tasks like summarizing financial or medical records. Few research systematically investigate methods for fine-tuning models to improve domain adaptation, while the majority of models are developed for general-purpose summarization. Furthermore, the absence of tools that make summarization understandable to non-technical users limits the models' practical application. Building upon this technical base, the project's objectives are to improve cutting-edge multilingual models for domain-specific activities and close the knowledge gap between research and practice by creating an intuitive, instructive web application.

4 Methods

To overcome the difficulties of multilingual document summary, this project uses a methodical approach to preprocessing, summarization, evaluation, and deployment. Strong input handling, including support for both text and PDF formats, is where the process starts. PyPDF2 is used to extract text from PDFs. To guarantee compatibility with subsequent operations, artifacts, invalid characters, and null bytes are eliminated during the cleaning process. The pipeline uses the langdetect package to implement language detection for non-English content. The Google Translator API is used to convert non-English text into English, and detected languages are mapped to English names for user feedback. The text is divided into digestible sections to handle long inputs and guarantee dependability, and retries are used to lessen translation errors.

Two cutting-edge transformer-based models, BART and PEGASUS, handle the summarization task. Using their pre-trained abilities, both models are optimized for document summarization and provide abstractive summaries. Using a pipeline approach, the BART model (facebook/bart-large-cnn) adds more heuristics for content augmentation, like extracting key phrases based on contextual patterns and frequency. In order to prioritize crucial information, it also assigns scores to phrases according to their positional, content, and length relevance. Longer texts that surpass BART's 1024 token input size restriction are split up into separate input chunks, each of which is summarized separately. The outputs are then combined into a final summary that is coherent. Similar tokenization and text processing techniques are used by PEGASUS (google/pegasus-large), which emphasizes abstractive summarization through beam search and length penalty adjustments. With the use of adjustable hyperparameters, both models can generate summaries that are short, medium, or long depending on the needs of the user.

The pipeline calculates a number of measures, such as ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and length ratio, to assess the caliber of the generated summaries. ROUGE metrics capture precision, recall, and F1 scores by measuring the n-gram overlap between the reference and the generated summary. BLEU compares the summaries to the reference at the token level to assess their fluency and sufficiency. An further viewpoint on how succinct the summaries are in comparison to the input material is offered by the length ratio. The nltk and rouge_score packages are used to calculate metrics, guaranteeing a consistent and repeatable assessment.

Comparative bar charts of important metrics for BART and PEGASUS are displayed by the program, which uses visualization tools with Plotly to offer insights into model performance. The Streamlit web application, which displays this comparison, has an easy-to-use tab-based interface that allows users to examine the original text, translations, summaries, and evaluation metrics. Additionally, users can download the data as Excel or CSV files for external examination, which include evaluation metrics and summaries.

The system is made to deal with issues that arise in the real world, like lengthy papers, loud inputs, and language hurdles. Segmentation techniques are used for lengthy texts, dividing the input into digestible chunks and analyzing each one separately. To guarantee cohesion and flow in the finished product, the summaries of these segments are further polished. Computational efficiency is increased by automatically allocating devices and enabling GPU utilization in contexts where it is accessible using PyTorch. The project bridges the gap between research and real-world implementation with this strong methodological framework, providing a user-friendly multilingual summarizing tool that supports a variety of use cases.

5 Strategy

The suggested multilingual document summarizing system uses a methodical, modular pipeline that maximizes usability, performance, and adaptability. The system generates high-quality, domain-specific English summaries from raw multilingual data by building on each step. Preprocessing, summary, and evaluation can be clearly implemented thanks to the modular design, which guarantees that each part is comprehensible on its own and works as a whole. The method is explained in three main steps below, followed by the integrated workflow in its entirety.

5.1 Step 1: Input Preprocessing and Language Standardization

At the most basic level, the first step is to transform incoming data into clean, language-standardized text that can be summarized, whether it is in the form of raw text or uploaded PDF files. The input's integrity and compliance with the summarization models are guaranteed by this preprocessing step.

- High-Level Overview: Text extraction and cleaning are the first steps in the pipeline. The

PyPDF2 library scans the contents of the document page by page in order to extract text if the input is a PDF. Cleaning is done on the extracted text to get rid of unnecessary spaces, invalid characters, and null bytes. The langdetect library is then used to identify the input's language, translating language codes into names that are understandable to the user. The Google Translator API is used to translate non-English input into English, chunking it to prevent problems for huge inputs.

- Detailed Description: For effective translation, preprocessing entails dividing the text into digestible parts (1000 characters). The translator receives each chunk and tries up to three times if it fails. This chunk-wise method prevents excessive memory consumption and guarantees robustness against connection faults.

5.2 Step 2: Abstractive Summarization Using BART and PEGASUS

The second stage uses two transformer-based models, BART and PEGASUS, to generate summaries once the input text has been preprocessed and standardized. This dual-model architecture supports a variety of summarizing requirements and allows for comparative performance study.

- High-Level Overview: The translated material is used by both models, which can provide short, medium, or long summaries. To create abstractive summaries, BART (Facebook/Bart-Large-CNN) employs a pipeline-based methodology that includes extra pre- and post-processing stages for content improvement. For huge inputs, PEGASUS (google/pegasus-large) uses sophisticated tokenization and chunking methods to conduct direct abstractive summarization.

- Detailed Description: Inputs that surpass their token limits—512 for PEGASUS and 1024 for BART—are split up into smaller parts for both models. A second summarization pass is performed for coherence after each segment has been individually summarized and the summaries are patched together. To make sure that important information is included, key terms that were retrieved during preprocessing are compared to the summaries. Beam size, repetition penalty, length penalty, and other hyperparameters are adjusted to strike a compromise between factual correctness and fluidity. The summarization flow is depicted in Figure 2, emphasizing the segment-wise processing and refining.

5.3 Step 3: Evaluation and Visualization

The final phase entails analyzing the summaries produced by both models, rating their performance and quality using a variety of measures. The user may make educated comparisons because these metrics are shown to them visually.

- High-Level Overview: Summaries are assessed using BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and length ratio in comparison to the original text. Bar charts are used to show the evaluation outcomes in order to make model comparison easier. To examine these findings in greater depth, the Streamlit program offers an intuitive user interface.

- Detailed Description: The evaluation module determines shortness by calculating length ratio, fluency by BLEU, and content preservation by n-gram overlap (ROUGE). Accurate and reliable metric computation is guaranteed by the rouge_score and nltk packages. Plotly is used to implement the visualization, allowing for interactive interaction with the evaluation findings. The metrics comparison is shown in Figure 3, which highlights how BART and PEGASUS perform differently in terms of fluency, precision, and recall.

5.4 Complete Algorithm

The last stage combines the elements into a coherent end-to-end algorithm, which is explained below:

Input Handling and Translation:

- Accept raw text or PDF as input.

- Extract text from PDFs and clean the content.

- Determine the input's language and use chunk-wise processing to convert non-English text to English.

Summarization:

- Tokenize and preprocess the cleaned, English-standardized text.
- Generate summaries using BART and PEGASUS, segmenting the text for long inputs.
- Post-process summaries by enhancing fluency and ensuring inclusion of key phrases.

Evaluation and Comparison:

- Compute evaluation metrics (ROUGE, BLEU, length ratio) for both summaries.
- Visualize metric scores in bar charts, enabling direct comparison of BART and PEGASUS outputs.

Output and Export:

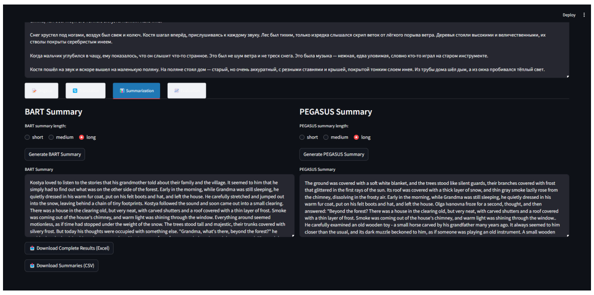
- Display summaries and evaluation results in the Streamlit interface.
- Offer downloadable files (CSV and Excel) containing summaries and metrics for external analysis.

While addressing real-world issues including multilingual input unpredictability and long-text summarizing restrictions, this modular approach guarantees smooth interaction between input, processing, evaluation, and user feedback while producing outputs of excellent quality.

6 Results

6.1 Comparative Evaluation of BART and PEGASUS

BART and PEGASUS's relative performance in terms of ROUGE and BLEU scores is depicted in the bar chart. In ROUGE measures, BART regularly fared better than PEGASUS, showing a greater degree of content overlap with the reference summary. PEGASUS did, however, show a modest advantage in BLEU scores, indicating that its summaries were more readable and fluid. These findings imply that PEGASUS provides outputs in a smoother, more natural language, but BART is superior at maintaining factual information.



Снег хрустел под ногами, воздух был свеж и колюч. Костя шагал вперёд, прислушиваясь к каждому звуку. Лес был тихим, только изредка слышался скрип веток от лёгкого порыва ветра. Деревья стояли высокими и величественными, их стволы покрыты серебристым инеем.

Когда мальчик углубился в чащу, ему показалось, что он слышит что-то странное. Это был не шум ветра и не треск снега. Это была музыка — нежная, едва уловимая, словно кто-то играл на старом инструменте.

Костя пошёл на звук и вскоре вышел на маленькую поляну. На поляне стоял дом — старый, но очень аккуратный, с резными ставнями и крышей, покрытой тонким слоем инея. Из трубы дома шёл дым, а из окна пробивался тёплый свет.

BART Summary

BART summary length:

○ short ○ medium ● long

BART Summary

Kostya loved to listen to the stories that his grandmother told about their family and the village. It seemed to him that he simply had to find out what was on the other side of the forest. Early in the morning, while Grandma was still sleeping, he quietly dressed in his warm fur coat, put on his felt boots and hat, and left the house. He carefully stretched and jumped out into the snow, leaving behind a chain of tiny footprints. Kostya followed the sound and soon came out into a small clearing. There was a house in the clearing old, but very neat, with carved shutters and a roof covered with a thin layer of frost. Smoke was coming out of the house's chimney, and warm light was shining through the window. Everything around seemed motionless, as if time had stopped under the weight of the snow. The trees stood tall and majestic, their trunks covered with silvery frost. But today his thoughts were occupied with something else. "Grandma, what's there, beyond the forest?" he

PEGASUS Summary

PEGASUS summary length:

○ short ○ medium ● long

PEGASUS Summary

The ground was covered with a soft white blanket, and the trees stood like silent guards, their branches covered with frost that glittered in the first rays of the sun. Its roof was covered with a thick layer of snow, and thin gray smoke lazily rose from the chimney, dissolving in the frosty air. Early in the morning, while Grandma was still sleeping, he quietly dressed in his warm fur coat, put on his felt boots and hat, and left the house. Olga Ivanovna froze for a second, thought, and then answered: "Beyond the forest? There was a house in the clearing old, but very neat, with carved shutters and a roof covered with a thin layer of frost. Smoke was coming out of the house's chimney, and warm light was shining through the window.. He carefully examined an old wooden toy - a small horse carved by his grandfather many years ago. It always seemed to him closer than the usual, and its dark muzzle beckoned to him, as if someone was playing an old instrument. A small wooden

Figure 1: Performance Comparison Chart 1

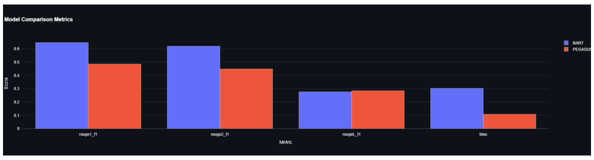
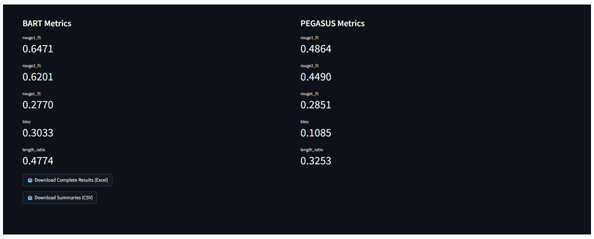


Figure 2: Performance Comparison Chart 2



BART Metrics	PEGASUS Metrics
rouge1_f1 0.6471	rouge1_f1 0.4864
rouge2_f1 0.6201	rouge2_f1 0.4490
rougeL_f1 0.2770	rougeL_f1 0.2851
bleu 0.3033	bleu 0.1085
length_ratio 0.4774	length_ratio 0.3253

Figure 3: Performance Comparison Chart 3

7 References

1. Scialom et al., MLSUM: The Multilingual Summarization Corpus, 2020
2. Ladhak et al., WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization, 2020
3. Liu et al., Multilingual Denoising Pre-training for Neural Machine Translation, 2020
4. Zhong et al., Attend to the Beginning: A Study on Using Bidirectional Attention for Extractive Summarization, 2020
5. Dou et al., Cross-Lingual Abstractive Summarization with Limited Parallel Resources, 2021
6. Lewis et al., MARGE: Pre-training via Paraphrasing, 2020
7. Conneau et al., XLM-RoBERTa: Unsupervised Cross-lingual Representation Learning at Scale, 2020
8. Shi et al., Neural Abstractive Text Summarization with Sequence-to-Sequence Models, 2019
9. Li et al., MultiSumm: Towards a Unified Model for Multi-Lingual Abstractive Summarization, 2021
10. Brown et al., Language Models are Few-Shot Learners, 2020
11. Dong et al., A Survey of Deep Learning Approaches to Text Summarization, 2021
12. Fan et al., Evaluating the Cross-Lingual Effectiveness of Massively Multilingual Neural Machine Translation, 2021
13. Chi et al., mT6: Multilingual Pretrained Text-to-Text Transformer with Translation Pairs, 2021
14. Kumar et al., Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation, 2022
15. Zhou et al., Cross-Lingual Summarization via Pre-Training, 2020
16. Tay et al., Efficient Transformers: A Survey, 2022
17. Chen et al., Understanding the Effectiveness of BERT in Multilingual Summarization, 2021
18. Wang et al., Domain Adaptation for Neural Abstractive Summarization, 2021
19. Liu and Liu, Evaluation Metrics for Text Summarization: A Survey, 2021