Summarization Model

Introduction

In a world overwhelmed by information, staying informed has never been more important—or more difficult. Every day, thousands of news articles are published in multiple languages across the globe. Yet, readers often lack the time to consume entire articles, especially when content is verbose, repetitive, or presented in a language they are less fluent in. This is where **automatic summarization** steps in as a poIrful tool.

The goal of this project is to develop a **multilingual news summarization system** that can generate clear, concise, and factually accurate summaries for articles written in **English, Spanish, and Mandarin Chinese**. The challenge is not just in summarizing text, but in preserving the core meaning and tone across languages with very different grammatical and cultural structures. Additionally, the system must generate summaries that are helpful to diverse readers while avoiding bias or misrepresentation.

To solve this, I build upon **transformer-based language models**, specifically leveraging the multilingual capabilities of **Facebook-bart**. I fine-tune this model on a curated multilingual dataset, ensuring it can understand and summarize news content across the three target languages. I evaluate the performance of the model using both automatic metrics like **ROUGE** and **BERTScore**, as well as through human-centered semantic analysis and multilingual bias detection.

Dataset & Preprocessing

To enable high-quality summarization across multiple languages, I curated and processed a multilingual dataset consisting of news articles in **English**, **Spanish**, and **Mandarin Chinese**. The dataset was constructed to reflect real-world journalistic content with sufficient lexical diversity, varying sentence structures, and culturally relevant framing.

Data Sources

The data was collected from multilingual news corpora that include both article bodies and corresponding human-written summaries (`xlsum`). These sources were selected based on three primary criteria:

1. **Content diversity** Coverage of politics, economy, health, technology, and culture.

2. **Language representation** Balanced samples from English, Spanish, and Chinese.

3. **Alignment** Availability of summaries that are contextually and semantically aligned with the source text.

Balancing Language Disparity

One of the key challenges in multilingual training is **language imbalance**. To mitigate this, I applied **downsampling** to normalize the number of samples per language to a common size of approximately **37,000 samples** each then downsampling to **10000 sample**s from each language for computational limits . This ensured that the model does not overfit to high-resource languages like English, while underperforming on lower-resource languages.

Preprocessing Pipeline

All text samples underwent the following preprocessing steps:

1. **Normalization** Unicode normalization, whitespace stripping, and punctuation standardization.

2. **Tokenization** Text was tokenized using the tokenizer associated with **`facebook/bart-large`**, which utilizes a byte-pair encoding (BPE) scheme optimized for English but extensible to multilingual text.

3. **Truncation and Padding** Inputs were truncated to a maximum of 512 tokens, and target summaries to 128 tokens. Padding was applied for batching purposes.

4. **Label Preparation** Labels were masked using the standard **`-100`** convention to ignore padded tokens during loss computation.

Model Architecture & Fine-Tuning Configuration

The backbone of this multilingual summarization system is the **`facebook/bart-large`** model a denoising autoencoder for pretraining sequence-to-sequence models. BART combines the bidirectional encoder of BERT with the autoregressive decoder of GPT, making it particularly well-suited for **abstractive summarization** tasks where fluency, structure, and generation quality are critical.

Given the computational intensity of **`facebook/bart-large`**, especially when handling long input sequences in a multilingual setting, several strategies were explored to optimize the training process under **hardware-constrained conditions**.

Baseline Fine-Tuning

Initial experiments were conducted using full fine-tuning of the **`facebook/bart-large`** model with a batch size of 8, mixed-precision training (FP16), and a learning rate of 1e-5. Generation was configured with beam search (num_beams=4) and a maximum output length of 128 tokens. The training used Hugging Face's **`Seq2SeqTrainer`**, with evaluation steps and checkpoints occurring every 100 steps.

Resource Constraints and Optimization Attempts

Despite optimizing the training arguments, fine-tuning **`bart-large`** remained **resource-intensive** (memory and time) on available hardware. To address this, a series of alternative approaches were explored:

1. Model Size Reduction

Smaller pre-trained summarization models such as **`facebook/bart-base`, `t5-small`, and `mbart50`** were trialed. However, the performance on multilingual summarization tasks particularly for Mandarin and Spanish was suboptimal due to limited pretraining exposure or reduced model capacity.

2. Quantization

To reduce memory footprint and speed up training, **8-bit quantization** was applied using **`bitsandbytes`**. While this allowed larger batch sizes to fit in memory, it came at the cost of **training instability and degraded performance**, especially in low-resource languages.

3. LoRA (Low-Rank Adaptation)

LoRA adapters were integrated using the **`peft`** library to perform parameter-efficient fine-tuning. This method aimed to fine-tune only a subset of low-rank weights while freezing the majority of the model. While this approach worked in terms of reducing GPU load, it also led to **higher loss values** and **lower ROUGE/BERTScore performance**, possibly due to underfitting in multilingual contexts.

Final Configuration

After extensive experimentation, the best trade-off between quality and resource feasibility was achieved by:

- Using **`facebook/bart-large`** with full fine-tuning
- FP16 mixed-precision training
- A reduced and balanced multilingual dataset (10k samples per language)
- Frequent evaluation checkpoints for early monitoring

Evaluation Methodology

To rigorously assess the quality and consistency of the multilingual summarization model, Iadopted a combination of **automatic metrics** and **manual semantic analysis**. These evaluations aimed to measure both surface-level textual overlap and deep semantic alignment between generated summaries and reference texts.

1. Automatic Metrics

I employed two complementary automatic evaluation metrics:

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

I used **ROUGE-1**, **ROUGE-2**, and **ROUGE-Lsum** to evaluate lexical overlap at different granularities:

- **ROUGE-1** measures unigram overlap (individual word recall).
- **ROUGE-2** captures bigram overlap (fluency and phrase matching).
- **ROUGE-Lsum** evaluates the longest common subsequence (structural alignment).

These metrics provide a traditional baseline for summarization quality, particularly useful for comparing performance across checkpoints.

1.2. BERTScore

To complement ROUGE, I used **BERTScore**, a neural semantic similarity metric based on contextual embeddings. BERTScore compares each token in the generated summary to its closest match in the reference summary using pre-trained multilingual transformers.

This method provides a better sense of **semantic fidelity**, especially in cases where the generated summaries use paraphrasing or differ in surface form.

Metric Configuration:

Check Notebook

2. Cross-Lingual Consistency Check

To assess the model's multilingual alignment, I compared semantically equivalent summaries generated from articles in different languages. I used BERTScore to compute pairwise similarity

3. Manual Human Evaluation

1. Accuracy
2. Coverage
3. Fluency & Readability
4. Tone Consistency

This analysis helped identify specific limitations such as:

- Occasional hallucinations in low-resource samples
- Slightly shorter output length than optimal in Mandarin
- Tone variability between English and Spanish summaries

4. Metric Progression Over Training

I tracked training and evaluation metrics over 10,000 steps to monitor learning progression. Loss decreased steadily, while ROUGE and BERTScore increased, indicating consistent improvement.

Metrics like **`ROUGE-L` and `BERTScore F1`** served as key indicators for early stopping and model selection

Conclusion

This project successfully developed a multilingual summarization system using the **`facebook/bart-large`** model, fine-tuned to generate concise and accurate summaries in English, Spanish, and Mandarin. Despite limited computational resources, various optimization strategies were explored—including quantization and LoRA—before settling on full fine-tuning for best performance.

The model was evaluated using ROUGE, BERTScore, and human analysis, showing strong semantic consistency and language-aware performance. By addressing challenges like language balancing and output quality, the system demonstrates the feasibility of building inclusive, cross-lingual NLP applications.

The final model is publicly available on Hugging Face and ready for further research or real-world deployment.