Enhancing Multi Target Cross-Lingual Summarization

Deviram Kondaveti
dk1273

Deepak Sai Are
da910

Sai Venkata Saketh Tunuguntla
st1269

Abstract

Cross-lingual summarization has emerged as a critical area of research in Natural Language Processing (NLP), aiming to generate concise and contextually accurate summaries in a target language from documents in a different source language. This study presents a novel methodology that significantly enhances cross-lingual summarization quality through the introduction of a beam-coherent summarization approach. Leveraging the multilingual transformer-based mBART model, our technique employs beam search coupled with semantic coherence reranking via sentence embeddings. Experimental results indicate that this approach effectively maintains coherence, contextual accuracy, and semantic integrity in multilingual summaries, demonstrating substantial improvements over existing methods.

1 Introduction

Cross-lingual summarization involves producing a summary in one language based on a source text written in another language. This task is essential in our increasingly globalized world, where information sharing across linguistic barriers has become a necessity for communication, education, research, and business. With the exponential growth of multilingual online content, there is a critical need for tools capable of distilling information efficiently and accurately, bridging gaps in language comprehension and accessibility.

Traditional summarization models often fail in cross-lingual scenarios because they primarily rely on language-specific training data and assumptions. These models struggle to maintain semantic coherence and accuracy when summarizing across languages, as translation involves intricate linguistic and cultural subtleties. Common problems include the misinterpretation of idiomatic expressions, cultural context losses, and degradation of factual accuracy during translation processes.

Recognizing these limitations, our project introduces an advanced cross-lingual summarization approach using a method we term "beam-coherent summarization." Our model leverages a powerful multilingual transformer model, which is adept at understanding and generating text across numerous languages. The distinctive element of our approach is its integration of beam search, an algorithmic strategy that simultaneously evaluates multiple potential summaries. To further enhance the quality of generated summaries, we implement semantic coherence reranking using sentence embeddings. This process ensures that the selected summaries are not only linguistically accurate but also contextually coherent, thus significantly improving their reliability and readability.

In addition to our beam-coherent method, several other prominent summarization methods exist. One such method involves direct multilingual translation models, which first translate the document into the target language and then perform summarization. Another method utilizes multilingual text-to-text transformer models, optimized specifically for summarization tasks across languages, providing direct summarization capabilities without explicit translation steps. Finally, methods based on large language models (LLMs) leverage extensive pre-training on diverse linguistic datasets, enabling them to perform summarization effectively due to their inherent understanding of language structure and semantics.

By addressing these challenges, our project aims to produce more reliable and contextually accurate summaries, thus substantially enhancing cross-lingual communication effectiveness. This research not only advances the field of NLP but also has the potential to impact practical applications across industries and academia by facilitating more effective global information dissemination.

2 Dataset

The primary dataset used in this research is the CrossSum dataset[1], a large-scale multilingual corpus constructed by aligning Wikipedia articles with corresponding Wikinews summaries across 45 languages. It encompasses over 1,600 language pairs, enabling both high-resource languages like French and Spanish and low-resource ones such as Amharic and Uzbek to be studied within the same framework. Each file in CrossSum is stored in `.jsonl` format, where every line contains a source document-summary pair, named as `<lang1>_<lang2>.jsonl`, where `lang1` is the source language and `lang2` is the target language—for example, `japanese_bengali.jsonl` contains a Japanese article and its Bengali summary.

Although originally designed for single source-target summarization, CrossSum required adaptation to serve the multi-target cross-lingual summarization goals of this study. To achieve this, we restructured the dataset by clustering multilingual documents that describe the same news story. These clusters were formed using graph-based maximal cliques, where each document is treated as a node and edges are drawn between nodes covering the same topic. Maximal cliques ensure that all documents in a cluster are mutually interconnected, forming a tight semantic group.

This clustering approach allows us to group up to seven document-summary pairs in different languages, treating each group as a unit for generating and evaluating multilingual summaries. As a result, we can assess not only the accuracy of individual language outputs but also the consistency of meaning across multiple generated summaries. This adaptation enhances the dataset's suitability for multi-target generation tasks and enables the development of more coherent and semantically aligned multilingual summarization systems. In this project, we focused on generating and evaluating summaries in six diverse target languages: English, Spanish, French, Portuguese, Russian, and Chinese (Simplified). These languages were selected to ensure a balance between high-resource and morphologically diverse languages.

[1] https://huggingface.co/datasets/csebuetnlp/CrossSum/blob/main/CrossSum.py

```
{
    "num_docs": int,
    "url0": str,
    "lang0": str,
    "text0": str,
    "summary0": str,
    "url1": str,
    "lang1": str,
    "text1": str,
    "summary1": str,
    ...
}
```

Figure 1: Each line corresponds to a cluster of documents and has the following format.

3 Multilingual Summarization Pipeline

3.1 Multilingual Transformer Backbone

mT5, a multilingual variant of the T5 model, serves as the core backbone of our system. Trained on the mC4 corpus containing over 100 languages, and fine-tuned on the CrossSum dataset, it is implemented via the Hugging Face Transformers library. mT5 follows a text-to-text framework where both input and output are in plain text, making it highly adaptable for multilingual summarization. This model delivers strong performance in high-resource languages like English, Spanish, and French. Its strengths include broad language coverage and consistent accuracy across major language pairs. However, in low-resource settings, performance can decline due to the scarcity of training examples.

3.2 Beam-Coherent Reranking

To enhance the coherence of summaries generated by mT5, we integrate a reranking component based on semantic similarity. Using the SentenceTransformers library and the all-MiniLM-L6-v2 variant of Sentence-BERT, we generate embeddings for the source document and beam-generated candidate summaries. These embeddings help us identify the candidate most semantically aligned with the input. This method is computationally lightweight and significantly improves semantic consistency, especially in multilingual outputs. Despite its efficiency, reranking does introduce minor latency, particularly in large-scale generation pipelines.

```
For the <source_lang> news article from BBC written below, provide a summary in <target_lang_1>, a summary in <target_lang_2>, ... and a summary in <target_lang_N>. All summaries should be one or two sentences long and follow the style of BBC. All summaries must contain the same information. Present the answer in the format of a JSON object where the keys are the language codes and the values are the summaries.

Text:

<source_document>
```

Figure 2: LLM Prompt.

3.3 Pivot-Based Translation

To expand language support, particularly for low-resource languages, we use a two-stage pipeline combining mT5 and Meta's No Language Left Behind (NLLB) model, accessed through Hugging Face Transformers. First, mT5 produces an English summary, which is then translated into the target language using NLLB. Reranking is applied before translation to ensure semantic coherence in the pivot English summary. This method is effective for preserving coherence even after translation, though translation artifacts or domain mismatches may occasionally degrade summary fluency or precision.

3.4 LLM-Based Summarization

For exploratory and flexible summarization, we utilize instruction-tuned Large Language Models such as Mistral-7B-Instruct and OpenAI's GPT-4o, accessed through the Hugging Face Transformers and OpenAI API libraries. The summarization process is guided by structured prompts that explicitly define the target language and stylistic requirements. These models excel at producing fluent and naturally phrased summaries, making them particularly suitable for scenarios where readability is a primary concern. Nevertheless, their lack of specialized fine-tuning for cross-lingual summarization tasks can lead to inconsistencies or overly verbose outputs, particularly across linguistically diverse inputs. Figure 2 illustrates the prompt formulation used to control the summarization behavior.

4 Evaluation Metrics

To evaluate the performance of our multilingual summarization system, we employed a diverse set of automatic metrics that capture both surface-level and semantic aspects of summary quality.

4.1 ROUGE

ROUGE is a lexical overlap metric that evaluates the similarity between generated and reference summaries based on n-gram matches, particularly unigrams and bigrams. It offers a quick and standardized method for performance comparison but is limited in capturing paraphrased or semantically equivalent content, especially in cross-lingual and abstractive summarization.

4.2 BLEURT

BLEURT employs a BERT-based pretrained model fine-tuned on human judgment data to assess summary quality. It evaluates fluency, relevance, and semantic adequacy, offering a closer approximation to human evaluation compared to traditional lexical metrics. Its ability to understand nuanced meaning makes it suitable for multilingual summarization tasks.

4.3 COMET

COMET is a reference-based evaluation metric tailored for cross-lingual applications. It utilizes multilingual encoder models to estimate the preservation of meaning across language boundaries. This makes it particularly effective for scenarios where the source and target texts are in different languages, such as in pivot-based summarization pipelines.

4.4 BERTScore

BERTScore uses contextual embeddings from transformer models to compute the semantic similarity between predicted and reference summaries. By evaluating cosine similarity at the token level, it captures paraphrases and deeper semantic relations, which are often missed by purely lexical approaches.

4.5 Language Accuracy

Language Accuracy checks whether the generated summary is in the intended target language. We utilize a fastText-based language identification model to verify each summary's output language, ensuring multilingual control and fidelity in generation.

5 Results

The results of our evaluation are shown in Figure 3 and Figure 4. In Figure 3, we compare the per-

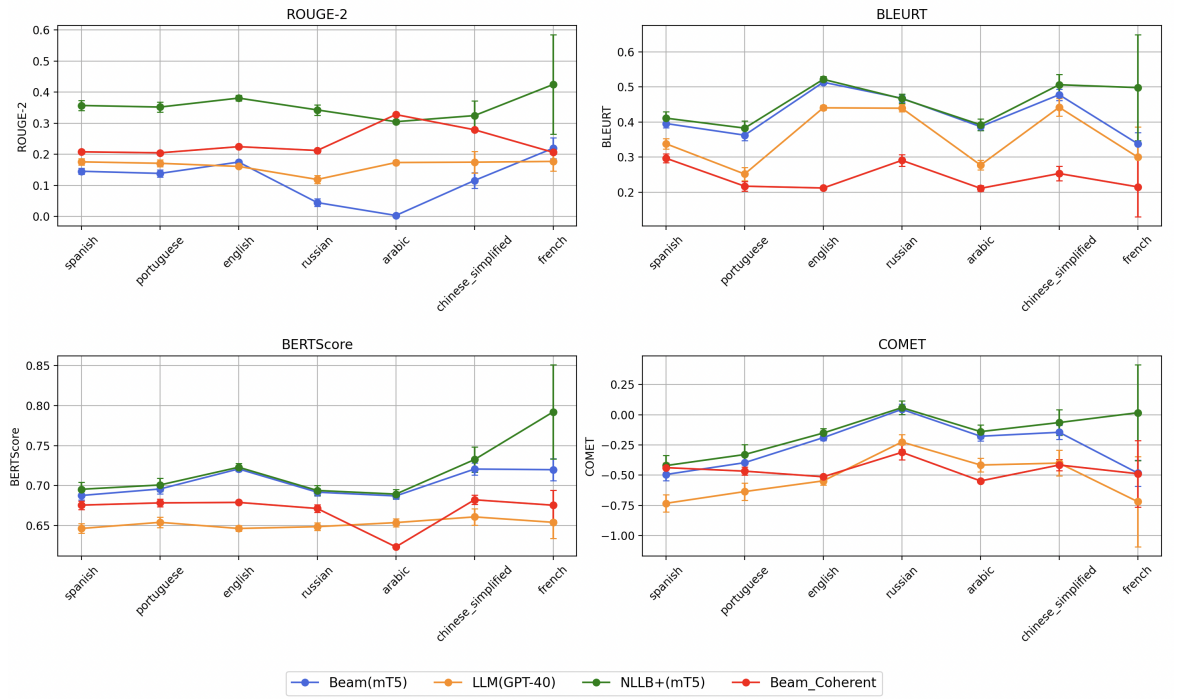

Figure 3: ROUGE-2, BLEURT, BERTScore and Comet metrics for 6 languages

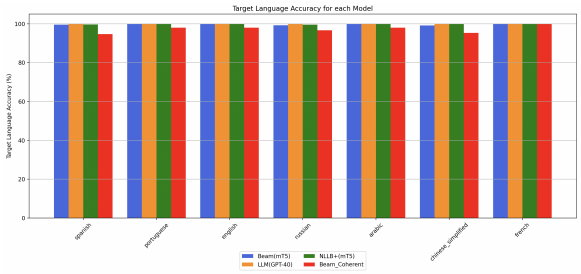

Figure 4: Target Language Accuracy for 6 languages.

formance of four different models Beam(mT5), LLM (GPT-4), NLLB+(mT5), and Beam Coherent using standard metrics like ROUGE-2, BLEURT, BERTScore, and COMET. Overall, the NLLB+(mT5) model performs best across most languages in all four metrics. It consistently scores higher in ROUGE-2, showing better word overlap with reference summaries, and achieves top scores in BLEURT and BERTScore, which indicate better semantic similarity. COMET scores also favor NLLB and mT5, reflecting its strength in producing fluent and accurate translations across languages. Beam Coherent performs slightly below NLLB+(mT5), but it maintains stable scores across all languages, which suggests that it helps improve consistency between summaries in different languages. The LLM(GPT-4) model has lower scores, especially in languages like Arabic and French.

Figure 4 shows the accuracy of each model in generating summaries in the correct target language. All models perform well, with scores close to 100%. for most languages. The NLLB+(mT5) model achieves perfect accuracy, while Beam(mT5) and LLM(GPT-4) also perform strongly. Beam Coherent shows slightly lower accuracy in a few languages, which may be due to the reranking process affecting the language output. Despite this, its overall performance remains reliable. These results suggest that NLLB+(mT5) is the most effective model both in quality and language accuracy, while Beam Coherent offers improved consistency across multiple target summaries.

Limitations

While the proposed framework demonstrates strong performance across multiple languages, it faces notable limitations. Summarization quality for low-resource languages remains uneven, with translation-based outputs sometimes suffering from fluency and accuracy issues. Pretrained models like mT5 and LLMs such as Mistral or GPT occasionally produce inconsistent or verbose summaries due to domain mismatch and lack of task-specific fine-tuning. The beam-coherent reranking strategy, although effective in improving semantic consistency, introduces additional computational overhead, lim-

iting its scalability. Furthermore, automated evaluation metrics like BLEURT and COMET, while advanced, may not fully capture deeper cultural or pragmatic nuances. Lastly, the dataset clustering process may introduce topic drift in some multilingual document groups, which can affect both training and evaluation reliability.

References

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7871–7880.

[2] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of NAACL*, pp. 483–498.

[3] Hasan, T., Bhattacharjee, A., Hasan, M. A., & Shahriyar, R. (2021). CrossSum: Beyond English-Centric Cross-Lingual Abstractive Text Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6001–6011.

[4] Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7881–7892.

[5] Rei, R., Farahani, M., Pavlick, E., Alves, R., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702.

[6] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*.