Natural Language Processing (CS 6120)

Literature Survey on RAG Based Multilingual News Retrieval

Group 7

Team Members

Anirudh Hegde (002813268)
Jahnavi Mishra (002724552)
Rahul Odedra (002835990)
Ronak Vadhaiya (002783524)

Submission Date

8th December, 2024

RAG Based Multilingual News Retrieval

Github Repository: *https://github.com/rahulodedra30/RAG-Based-Multilingual-News-Retrieval/tree/main*

Dataset: *https://huggingface.co/datasets/reciTAL/mlsum*

1.Abstract

The **RAG-Based Multilingual News Retrieval System** addresses the challenge of accessing and summarizing multilingual news content. The system integrates Retrieval-Augmented Generation (RAG) to enable cross-lingual news retrieval, translation, and summarization, all in one streamlined process. Users submit queries in English and receive summaries of the top 5 most relevant news articles drawn from a knowledge base of multilingual articles. Key technologies include MBART50 for translation, SBERT embeddings for similarity, and FAISS for fast retrieval. This project builds on the RAG model proposed by Lewis et al. (2020), with enhancements to support multilingual retrieval. The system is evaluated using BLEU, BERTScore, Cosine Similarity, and SBERT Score, ensuring high relevance, quality, and user satisfaction. The interactive Streamlit-based UI offers an accessible, efficient, and human-centered user experience.

2.Introduction

Background

In a globalized world, access to multilingual news is essential for gaining a complete perspective on important issues. However, traditional search engines and news platforms are often monolingual, focusing on English-language content. This restricts users' access to crucial content from non-dominant languages, causing an information gap. Even when content in other languages is available, users are forced to rely on translation tools or separate searches in different languages, resulting in fragmented, inefficient information retrieval.

To address these challenges, we propose a RAG-Based Multilingual News Retrieval System. This system enables users to submit an English query and receive context-aware

summaries of multilingual news articles. By integrating RAG (Retrieval-Augmented Generation), the system incorporates retrieval, translation, and summarization into one streamlined process. This approach offers users access to a more diverse, multilingual perspective on world events.

Objective

The primary objectives of this project are:

1. Enable users to search for multilingual news using an English query.
2. Retrieve the top 5 most relevant news articles from a large knowledge base.
3. Summarize the retrieved articles into concise English summaries.
4. Provide an interactive, user-friendly UI for query input and result exploration.

Scope

This project works with multilingual news articles from **German, Spanish, French, Russian, Turkish, and Arabic**. It integrates the following key components:

1. **RAG Model**: Combines retrieval and generation to produce context-aware summaries.
2. **Multilingual Knowledge Base**: Built using a FAISS index of translated news articles.
3. **Translation System**: Uses MBART50 to translate non-English articles into English.
4. **Summarization**: Uses MBART50 to generate concise, clear, and human-readable summaries.
5. **Evaluation Metrics**: Ensures high-quality results using BLEU, BERTScore, Cosine Similarity, and SBERT Score.
6. **Interactive UI**: Powered by Streamlit, allowing users to submit queries, review results, and explore related articles.

3.Methodology

Search Strategy

1. **Data Sources**:
    - MLSUM Dataset (HuggingFace) for German, Spanish, French, Russian, and Turkish articles.
2. **Search Engines**:
    - HuggingFace Datasets API for large-scale multilingual news datasets.
    - Kaggle API for collecting Arabic news articles.
3. **Search Criteria**:
    - Collect a minimum of 700 articles per language.
    - Filter for articles with clear title, content, and summary.

Keywords

Key terms for data collection and model development include:

- **"Multilingual News Retrieval"**
- **"Cross-Lingual Summarization"**
- **"RAG for Cross-Lingual Search"**
- **"Dense Retrieval and Embeddings"**

Selection Criteria

- **Languages**: Include news articles from German, Spanish, French, Russian, Turkish, and Arabic.
- **Relevance**: Articles must contain clear titles, summaries, and main content.

- **Completeness**: Only articles with complete and well-structured summaries are included.

4.Literature Review

Thematic Analysis

Theme 1: Retrieval-Augmented Generation (RAG) for Knowledge-Intensive Tasks

Reference: Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

- **Key Findings**:
    - Introduced **RAG**, which combines retrieval and generation for knowledge-driven tasks.
    - Demonstrated the power of **combining retrieval with generation** in open-domain question-answering.
    - Highlighted RAG's ability to retrieve from a large knowledge base on demand.

Theme 2: Multilingual Summarization

Reference: Ladhak, F., et al. (2020). WikiLingua: A Multilingual Abstractive Summarization Dataset.

- **Key Findings**:
    - Presented the **WikiLingua dataset**, a large multilingual dataset for cross-lingual summarization.
    - Demonstrated that models like **MBART50** can be fine-tuned on multilingual datasets for better summarization performance.
    - Introduced **cross-lingual transfer learning** as a strategy for summarizing low-resource languages.

Theme 3: Dense Passage Retrieval (DPR) for Information Retrieval

Reference: Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering.

- **Key Findings**:
    - Introduced **Dense Passage Retrieval (DPR)**, which uses dense embeddings for passage retrieval.
    - Showed that **dense embeddings outperform BM25** in large-scale search systems.
    - Provided a framework for **efficient large-scale retrieval using FAISS indices**.

Comparative Analysis

1.Methodologies

Key Insights

- **Lewis et al. (2020) -** RAG: Combines retrieval and generation into a unified framework. It retrieves documents from a large knowledge base and uses BART to generate a natural language response.
- **Ladhak et al. (2020) -** WikiLingua: Focuses on multilingual abstractive summarization for 18 languages. The system uses mBART50 to generate cross-lingual summaries, but it lacks a retrieval component.
- **Karpukhin et al. (2020) -** DPR: Introduces Dense Passage Retrieval (DPR) for large-scale passage retrieval. It builds a FAISS-based knowledge base to store dense embeddings for fast, accurate search. Unlike RAG, DPR does not generate responses but instead retrieves passages.

2. Datasets

Key Insights

- **Lewis et al. (2020) -** RAG: Relies on English-based QA datasets like Natural Questions and TriviaQA. It uses Wikipedia as its external knowledge source, from which relevant documents are retrieved.
- **Ladhak et al. (2020) -** WikiLingua: Focuses on summarizing multilingual content using document-summary pairs. It has data for 18 languages, making it highly valuable for multilingual summarization research.
- **Karpukhin et al. (2020) -** DPR: Uses Wikipedia passages as a dense knowledge base. The model retrieves passages to support downstream systems like RAG for knowledge generation.

3.Results

Key Insights

- **Lewis et al. (2020) -** RAG: Improved over standard BART, achieving a +4% gain in Exact Match (EM) and F1 scores.
- **Ladhak et al. (2020) -** WikiLingua: Achieved strong performance on ROUGE scores, surpassing other summarization baselines.
- **Karpukhin et al. (2020) -** DPR: Demonstrated a +5% improvement over BM25 on Top-K retrieval accuracy, making it the standard model for dense passage retrieval.

4.Strengths and Weaknesses

Paper	Strengths	Weaknesses
Lewis et al. (2020) - RAG	Combines retrieval and generation into one system	Limited to English only; computationally expensive
Ladhak et al. (2020) - WikiLingua	Multilingual summarization for 18 languages	No retrieval system; limited to document-summary tasks
Karpukhin et al. (2020) - DPR	Dense embeddings outperform BM25	No summarization; only supports English

Key Insights

- **Lewis et al. (2020) -** RAG: Excellent at unifying retrieval and generation, but is limited to English-language queries.
- **Ladhak et al. (2020) -** WikiLingua: Supports multilingual summarization but lacks a retrieval mechanism.
- **Karpukhin et al. (2020) -** DPR: Sets a new benchmark for dense retrieval but does not support generation or summarization.

Aspect	Lewis et al. (2020) - RAG	Ladhak et al. (2020) - WikiLingua	Karpukhin et al. (2020) - DPR
Primary Focus	Unified RAG system (retrieval + generation)	Multilingual summarization	Dense passage retrieval
Language Support	English only	18 languages (multilingual)	English only
Core Contribution	RAG system for QA and knowledge-intensive tasks	Multilingual summarization dataset (WikiLingua)	DPR model for dense retrieval
Main Limitation	No multilingual support	No retrieval	No summarization, limited to English

Conclusions from Comparative Analysis

The three papers analyzed highlight key building blocks for **RAG-Based Multilingual News** Retrieval Systems. By combining the RAG system (Lewis et al., 2020), multilingual summarization (Ladhak et al., 2020), and dense retrieval with FAISS (Karpukhin et al., 2020), a comprehensive RAG system can be developed. The following insights can be drawn:

- Lewis et al. provides a framework for retrieval + generation, but it is limited to English.
- Ladhak et al. demonstrates multilingual summarization capabilities, but it lacks a retrieval component.
- Karpukhin et al. focuses on dense retrieval but does not include summarization.

A hybrid system could be created by combining the strengths of all three approaches:

1. Use dense retrieval (DPR) to locate relevant multilingual articles.

2. Translate them into English using MBART50.
3. Summarize them using RAG's generator or mBART50.

5.Critical Analysis

Evaluation of Research Quality

The quality of the selected studies is evaluated based on the validity, reliability, impact, and significance of their findings. Each paper's contribution to the development of the RAG-Based Multilingual News Retrieval System is assessed as follows:

1. Lewis, P., et al. (2020) - Retrieval-Augmented Generation (RAG) for Knowledge-Intensive NLP Tasks

- **Validity and Reliability**:
    - The RAG model has been extensively tested on widely recognized datasets such as Natural Questions (NQ), WebQuestions, and TriviaQA. The use of established datasets enhances the reliability and reproducibility of the results.
    - The research employs strong **evaluation metrics** (Exact Match, F1 Score) to assess performance, and the results have been validated by multiple independent researchers, highlighting the robustness of the proposed method.
- **Impact and Significance**:
    - RAG introduces a groundbreaking approach by integrating retrieval and generation in a single framework. This innovation sparked new research into knowledge-augmented models for information retrieval and question-answering systems.
    - The significance of the RAG model is seen in its impact on the development of **ChatGPT, Bing Chat, and Google Bard**, as these systems now incorporate retrieval-based approaches for open-domain question-answering.

2. **Ladhak, F., et al. (2020) - WikiLingua: A Multilingual Abstractive Summarization Dataset**

- **Validity and Reliability**:
    - The study introduces the WikiLingua dataset, which supports multilingual abstractive summarization. This dataset is publicly available and widely adopted in the NLP community, promoting transparency and reproducibility.
    - Their empirical evaluations include diverse summarization models like mBART, BERTSum, and PEGASUS, ensuring that the results are validated on multiple architectures.
- **Impact and Significance**:
    - The WikiLingua dataset fills a major gap in the field of multilingual summarization by providing paired datasets for over 18 languages.
    - This research enables advances in cross-lingual learning and enhances performance for low-resource languages. It provides a crucial resource for projects like multilingual summarization and cross-lingual summarization.

3. **Karpukhin, V., et al. (2020) - Dense Passage Retrieval (DPR) for Open-Domain Question Answering**

- **Validity and Reliability**:
    - The study proposes **Dense Passage Retrieval (DPR)**, which outperforms traditional retrieval methods like **BM25** on multiple benchmarks.
    - DPR's results are supported by a thorough comparative analysis with existing models, ensuring that the performance improvement is valid and statistically significant. The model's effectiveness is further validated through its integration with the **FAISS index**, which has since become the standard for dense retrieval.
- **Impact and Significance**:

- DPR's impact extends to search engines, Q&A systems, and RAG-based retrieval models.
- By shifting from sparse BM25 search to dense vector-based search, it has significantly improved retrieval precision and recall, which is essential for large-scale retrieval tasks. This approach has been widely adopted in systems like Google Search and ChatGPT.

Identification of Gaps

Despite the strengths of the selected research, several key **gaps and areas for future investigation** remain:

1. Lack of Cross-Lingual Retrieval in RAG Models

- **Problem**: Current RAG models (e.g., Lewis et al., 2020) focus on English-only retrieval. Cross-lingual retrieval is not explicitly addressed, leaving a gap in multilingual content accessibility.
- **Impact**: Users may miss out on relevant information available in non-English articles.
- **Future Work**: Develop RAG models that handle cross-lingual queries by embedding multilingual context and supporting multilingual FAISS indices.

2. Limited Dataset Diversity

- **Problem**: The datasets used by Lewis et al. and Karpukhin et al. are predominantly focused on English-language datasets like Natural Questions (NQ) and WebQuestions.
- **Impact**: This limits the generalizability of retrieval models when applied to non-English articles.
- **Future Work**: Include datasets from underrepresented languages like Hindi, Swahili, and Indigenous languages. This can be done using datasets like WikiLingua and MLSUM, which offer multilingual articles.

3. Efficiency and Scalability

- **Problem**: FAISS-based dense indices grow significantly as the knowledge base size increases. Storing and searching large indexes in low-memory environments becomes challenging.
- **Impact**: Resource constraints could make it difficult to use RAG systems for large-scale multilingual retrieval.
- **Future Work**: Use HNSW (Hierarchical Navigable Small World Graphs) to reduce memory consumption and enhance retrieval speed.

4. Unified Retrieval-Translation-Summarization Pipeline

- **Problem**: Most RAG implementations work in a single language. There is no unified pipeline to retrieve multilingual content, translate it, and summarize it in English.
- **Impact**: Current models require three separate components (retriever, translator, summarizer) instead of a unified, end-to-end system.
- **Future Work**: Train an end-to-end model that supports retrieval, translation, and summarization using a single RAG pipeline.

Implications

Practical Implications

1. **Global News Access**: Users can access regional news perspectives from non-English-speaking regions, reducing reliance on monolingual news sources.
2. **Multilingual Chatbots**: RAG-based multilingual models can be integrated into conversational AI systems like ChatGPT and customer support systems.
3. **Cross-Lingual Knowledge Management**: Organizations can access global insights from various languages, allowing them to make informed, data-driven decisions.

Theoretical Implications

1. **Unified RAG Pipeline**: The development of an end-to-end retrieval, translation, and summarization system will establish a new paradigm for multilingual RAG models.
2. **Low-Resource Language Inclusion**: The development of cross-lingual models will bring equity in access to information for low-resource languages.
3. **Multilingual Embeddings**: Advances in multilingual embedding models like SBERT support large-scale multilingual content retrieval and summarization.

Limitations

1. Language Support

- The selected studies focus on European languages (German, French, etc.) but lack support for **underrepresented languages** (e.g., Swahili, Hindi, Bengali).

2. Computational Constraints

- Memory limitations affect the use of large FAISS indices for dense retrieval. As the index size grows, so does the memory footprint.
- Future models should adopt techniques like approximate nearest neighbor search (ANN) and HNSW to reduce computational complexity.

3. Limited Cross-Lingual Context

- **Problem**: Retrieval systems do not incorporate cross-lingual context in embeddings, meaning embeddings for the same query in English, German, and Spanish may differ.
- **Solution**: Train models to contextually align multilingual embeddings so that similar queries produce similar embeddings regardless of the language.

6.Conclusion

Summary of Findings

This report evaluated three critical research papers on **RAG-based retrieval, summarization, and dense passage retrieval**. Key takeaways include:

1. RAG Models (Lewis et al., 2020) highlight the effectiveness of combining retrieval and generation.
2. WikiLingua (Ladhak et al., 2020) introduces the WikiLingua dataset, which supports multilingual summarization.
3. Dense Retrieval (DPR) (Karpukhin et al., 2020) introduces FAISS-based retrieval using dense embeddings, which are critical for our RAG system.

Future Directions

1. **Cross-Lingual Retrieval**: Develop **RAG models that support multilingual queries** and retrieval from multilingual sources.
2. **Multilingual Datasets**: Expand datasets to include **low-resource languages**.
3. **Optimized FAISS Indices**: Use **HNSW (Hierarchical Navigable Small World)** graphs to reduce storage costs and improve retrieval efficiency.
4. **Unified RAG Pipeline**: Integrate **retrieval, translation, and summarization** into a single end-to-end RAG model.

7.References

1. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
2. Ladhak, F., et al. (2020). WikiLingua: A Multilingual Abstractive Summarization Dataset.
3. Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering.

8.Group Reflection

The development of the RAG-Based Multilingual News Retrieval System was a valuable learning experience for our team. We gained hands-on experience with retrieval, translation, summarization, and user interface development. This project enhanced our understanding of RAG, SBERT embeddings, FAISS indexing, and multilingual NLP.

Challenges Faced:

- Memory Constraints: Managing large FAISS indices required optimization to handle large datasets.
- Language Alignment: Ensuring consistent translations for multiple languages using MBART50 was challenging.
- System Integration: Combining retrieval, summarization, and translation into a unified workflow required effective collaboration.

Task Distribution:

- **Rahul**: Responsible for data processing and translation using MBART50.
- **Anirudh & Jahnavi**: Worked on summarization and RAG development, focusing on retrieval, generation, and FAISS-based search.
- **Ronak**: Developed the Streamlit user interface (UI), making the system interactive and user-friendly.

Through teamwork and regular sync-ups, we tackled these challenges and built a comprehensive system. The project strengthened our problem-solving, technical, and collaborative skills, and it provided a deep understanding of cross-lingual retrieval systems.