Natural Language Processing

Project Proposal

RAG Based Multilingual News Retrieval

Group 7

RAG Based Multilingual News Retrieval

Basic Idea

In an era of information overload, users face significant challenges when trying to find relevant and diverse news articles across multiple languages. Existing retrieval systems are often limited to monolingual or regional sources, leading to bias in available information and neglect of pertinent content in non-dominant languages. Users frequently encounter the frustration of missing critical news perspectives or context, especially when tracking global issues or looking for regional insights in a preferred language.

Even if the user attempts to retrieve articles in multiple languages, current solutions often require separate, language-specific searches, leading to inefficient information retrieval and fragmented results. This lack of an integrated, multilingual retrieval system results in an incomplete, often one-sided view of events.

Approach to Solution

To address this, we propose a custom Retrieval-Augmented Generation (RAG) model that will allow users to search for relevant news articles across multiple languages based on an input query, returning summarized content in the English Language. The RAG model will leverage a multilingual knowledge base of news articles, integrating state-of-the-art language models for both retrieval, summarization and generation tasks. This setup enables cross-lingual search and translation, ensuring that relevant content is accessed from multiple regions and languages.

The workflow is as follows:

1. **Multilingual News Articles Collection:**

    - **Data Source**: Using Wiki-Lingual Hugging Face dataset and Arabic News Articles Dataset from Kaggle, we collect news articles from various domains and languages.

        - Dataset: https://huggingface.co/datasets/esdurmus/wiki_lingua
        - Dataset:https://www.kaggle.com/datasets/haithemhermessi/sanad-dataset/data

    - **Domain-Specific Filters**: For optimized retrieval and relevance, articles are pre-filtered by specific domains, e.g., science, business, or regional news, which can be adjusted based on user needs.

2. **Translation and Summarization Pipeline:**

   - **Summarization Model:** A transformer-based summarization model like BART or T5, fine-tuned on news article datasets with summaries in English, will condense the retrieved articles. This model provides concise summaries, highlighting key points for quick consumption.

   - **Translation Model for Consistency:** For non-English articles, a model like mBART performs the translation to English. This translation step ensures that all summaries are consistently in English, regardless of the original language.

   - **Summarization Fine-Tuning:** The summarization model is trained on a dataset with paired articles and summaries to ensure clarity and relevance in news summaries.

3. **Multilingual Transformer-Based Retriever:**

   - **Creating the Knowledge Base:** Articles from News dataset is translated to English using a translation model (e.g., mBART for multilingual translation).

   - **Embedding Generation and Storage:** Each translated article is embedded into a vector using multilingual transformer models like 'mBERT' to capture cross-lingual semantics. These embeddings are stored in a vector database, enabling fast retrieval based on user queries.

   - **Efficient Retrieval:** When a user submits a query, the system uses the multilingual retriever to identify the top n most relevant articles, ensuring the retrieval is language-agnostic and reflects diverse perspectives.

4. **User Query Processing and Output Generation:**

   - **User Input:** The user inputs a query on a specific topic or event, which is converted into an embedding using a transformer model.

   - **Retrieve Top N Articles:** The vector database is queried to retrieve the N most relevant articles based on embedding similarity to the query.

   - **Summarized Output:** The system returns summaries for the top N articles, each capturing key details from the original articles, providing a brief, coherent response aligned with the user's query.

Why It’s Useful

This tool lets users follow international news more easily, especially from regions where English news coverage might be limited. By combining translation, relevant article retrieval, and summarization, the system provides an efficient, accessible way to stay informed about the world, all tailored to each user’s interests.

Related Work

1. Lewis et al., 2020, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks” – This paper presents the foundational RAG model and explores the application of RAG to improve answer generation by retrieving relevant documents. The work demonstrates that retrieval-augmented methods outperform traditional models in knowledge-intensive tasks by leveraging both retrieval and generation capabilities.

   2. Ladhak et al., 2020, “WikiLingua: A Multilingual Abstractive Summarization Dataset” – This research explores cross-lingual summarization, presenting a large-scale multilingual dataset. The findings underscore the importance of multilingual models for effective cross-lingual summarization, and the work provides insights into handling language diversity and translation for summarization tasks.

Assessment Methodology

1. Performance Evaluation Metrics:

- BLEU Score and ROUGE-L for translation and summarization accuracy, ensuring meaningful and precise translations.
- Mean Average Precision (MAP) for retrieval relevance, to ensure user-specific topics are accurately matched with the most pertinent articles.
- User Relevance Feedback: We will measure user satisfaction via surveys or feedback scores, assessing the relevance and readability of the retrieved content.

2. Cross-Validation Strategy:

- **K-Fold Cross-Validation**: Applied to the multilingual retrieval and summarization pipeline, ensuring model robustness and identifying optimal hyperparameters across various languages.

- **Leave-One-Out Cross-Validation**: Used specifically in fine-tuning the generative summarization model to handle different language inputs accurately and efficiently.

3. Ablation Settings:

- **Data Input Variations**: We will experiment with datasets of differing sizes and language complexities to understand their impact on retrieval and summarization.
- **Algorithm Complexity**: Testing simpler and more complex RAG architectures, such as various transformer layers and embeddings, to determine the best trade-off between performance and processing time.
- **Preprocessing Techniques**: Assessing the impact of various tokenization, language normalization, and translation preprocessing steps on the system's output quality.

Ablation experiments will help rank each component's impact on final performance, guiding us in optimizing the model for accurate and efficient information retrieval.