CRISISFacts: Real-Time Crisis Summarization

A System for Efficient Summarization of Crisis-related Information

Isfar Baset, Ziyan Di, Sheeba Moghal, Bella Shi and Jacky Zhang

Introduction

The project addresses a theme commonly discussed in class, “we are drowning in information, yet striving for knowledge” by building a hybrid text retrieval and summarization system to efficiently query and summarize event-based datasets. By combining **BM25** for precise keyword-based retrieval and **dense embeddings** for semantic understanding, the system ensures both relevance and accuracy in information retrieval. It offers two modes of interaction:

1. **Event and Class Label Summarization**: Users specify events and labels for targeted summaries.
2. **Custom Query Summarization**: Free-text queries allow contextual searches and automated summarization.

This approach bridges the gap between the overwhelming volume of data and the need for meaningful, actionable insights.

Methods

1. Data Collection

- Sources: The dataset used for this project is the **TREC 2023 CrisisFACTS Track** Dataset, published by the National Institute of Standards and Technology (NIST). This dataset focuses on temporal summarization for first responders in emergency situations and provides crisis-related information in short, time-ordered updates.
- Dataset: The structure of the dataset contains textual records related to crisis events (e.g., earthquakes, hurricanes, floods). Includes **metadata** fields such as:
    - **Event**: The type or name of the crisis (e.g., "2015 Nepal Earthquake").
    - **Class Label**: Specific information categories (e.g., "infrastructure damage," "donation requests").
    - **Cleaned Text**: Short textual descriptions of updates or crisis-related facts.

2. Data Preprocessing

- Text Cleaning:
    - **Removal of Noise:** URLs, mentions (@usernames), and retweet markers (RT) are removed to eliminate irrelevant tokens. Special characters and emojis are stripped to standardize the text.
    - **Case Normalization**: All text is converted to lowercase to ensure uniformity and facilitate retrieval.

    - **Whitespace Handling:** Extra spaces and line breaks are condensed for cleaner inputs.
- Language Filtering: Non-English records are filtered out to focus the analysis on English texts, ensuring compatibility with the summarization model.
- Normalization:
    - **Query Normalization:** User queries are transformed into a standardized format (e.g., replacing spaces with underscores, removing filler phrases). This ensures consistent matching with event and class label names in the dataset.

3. Data Retrieval

- Approach: Combines **BM25** (keyword-based retrieval) and **dense embeddings** (semantic similarity) for accurate and contextually relevant results.
- Process:
    - **Step 1: BM25 Scoring**: Queries are tokenized, and BM25 calculates relevance scores based on keyword matches.
    - **Step 2: Dense Embedding Scoring**: Queries are embedded using a pre-trained DistilBERT model. Cosine similarity measures the semantic closeness of these embeddings to document embeddings

4. Summarization

- **Models Used:**
    i. **BART Large CNN**: Chosen for its extractive summarization capabilities and ability to handle large input sequences effectively.
- **Input:**

    A concatenated cleaned_text of rows filtered through hybrid retrieval.
- **Output:**
    A concise, human-readable summary relevant to the query or selected event and label.

5. User Interface

- Frontend:
    i. Built using Flask and Jinja templates.
    ii. Designed for user-friendly interaction with two modes:
        1. **Event and Class Label Summarization**.

2. **Custom Query Summarization**.

- Features:
  i. **Event Selection**: Displays human-readable event names for intuitive interaction.
  ii. **Class Label Selection**: Dynamically retrieves available labels for chosen events.
  iii. **Summarization Display**: Generates summaries in real time based on user input.
  iv. **Custom Query Support**: Allows free-text input for searching across events.

6. Evaluation

- The system was evaluated using representative queries:

  Event-Specific Query:
  Example: “2015 Nepal Earthquake” → Generates summaries for specific events and class labels.

  Custom Query:
  Example: “Find updates on donations for the Nepal Earthquake” → Processes the query and retrieves top results for summarization.

Key Findings

Strengths of the Approach

The hybrid retrieval and summarization system combines the precision of **BM25** for keyword-based matching with the semantic understanding of **dense embeddings**, ensuring both accuracy and contextual relevance. By integrating query normalization and human-readable transformations, the system enhances user experience, making event and class label selection intuitive and accessible. The use of a pre-trained summarization model, such as **BART**, allows for concise and meaningful outputs, preserving essential information from large datasets. This approach ensures efficiency, flexibility, and usability, making it well-suited for real-world applications across diverse user groups.

Challenges and Limitations

- **Embedding Computation**: Precomputing dense embeddings for large datasets is resource-intensive.
- **Summarization Model Limitations**: Text truncation might miss some relevant information in longer inputs.
- **Scalability**: Processing very large datasets may result in high CPU and memory usage.

Real-World Applications

This hybrid text retrieval and summarization system has significant real-world applications across various domains. In **crisis management systems**, it can efficiently summarize and retrieve critical updates during natural disasters, helping emergency responders and government agencies stay informed in real-time. For **humanitarian aid organizations**, the system can assist NGOs and relief agencies by identifying key needs such as donation requests, infrastructure damage or affected individuals from vast event-based datasets. Media organizations can leverage the system for **news monitoring**, extracting concise and relevant information from large volumes of event data to enhance public communication. Additionally, the system can provide a valuable service for **international students and expatriates** living far from their home countries. By generating real-time summaries of ongoing crises such as natural disasters, political unrest or humanitarian events, it can help individuals stay updated and connected to the situation in their home regions, reducing anxiety and enabling timely responses. The system can also be adapted for **academic research** and **policy analysis**, where summarizing and retrieving data from large repositories is essential for decision-making and reporting. Its ability to combine precision and semantic understanding makes it versatile for use cases requiring both keyword-based accuracy and contextual relevance.

Potential Improvement & Future Work

Several enhancements can improve the system's efficiency and scalability. Optimizing **embedding computation** through distributed computing or caching mechanisms would reduce resource overhead, especially for large datasets. For scalability, integrating with tools like **Elasticsearch** could enable faster hybrid retrieval and improved performance on massive datasets. In terms of summarization, experimenting with advanced transformer models such as **T5** or **Longformer** can help handle longer input sequences more effectively and improve summary quality. Future iterations could include **interactive visualizations**, such as dynamic charts or word clouds, to provide more insightful representations of summarized content. Adding **multilingual support** would also broaden the system's applicability, allowing it to process and summarize data in non-English languages. By addressing these areas, the system can evolve into a more robust, efficient, and user-friendly solution capable of serving diverse global needs.

Conclusion

This project successfully demonstrated a hybrid retrieval and summarization system for event-based datasets. By combining BM25 and dense embeddings, the system ensures precise and contextually relevant results. It offers an intuitive user experience, supports targeted and custom query summaries, and addresses real-world challenges in information retrieval. Future improvements will focus on scalability, advanced summarization techniques, and broader dataset support.

References

1. National Institute of Standards and Technology. (2024, September 11). *TREC 2023 CrisisFACTS Track Dataset* [Dataset]. Data.gov. https://catalog.data.gov/dataset/trec-2023-crisisfacts-track-dataset