CrisisFacts Summarization Using Multi-Stream Data for Emergency Management

Team-14 Rescue-Retrievals

Abstract

Effectively summarizing crisis data is crucial for providing emergency responders with timely and actionable insights. This study helps to understand the challenges of processing noisy and diverse multi-stream datasets by using a structured pipeline combined with semantic and summarization methods technique. We integrate BM25-based retrieval with BERT-enhanced relevance scoring to create summaries to specific queries with key facts. The proposed approach performs competitively with top scores from the TREC CrisisFACTS event while uncovering challenges in handling diverse data sources and maintaining semantic accuracy. Additionally, the dual approach of extractive and abstractive summarization addresses both real-time and contextual needs, emphasizing the value of semantic alignment for generating high-quality summaries.

Keywords

Extractive Summrization, Abstractive Summrization, BART Model, BERT Model, Multi-streaming Data, Fact Finding, Retrieval, Evaluation,

1. Problem Statement

In crisis situations, emergency responders need quick, accurate, and clear information to make informed decisions. Given a set of queries representing specific needs and organized multi-stream data from sources like social media, news articles, and other feeds, the task is to create concise daily summaries. These summaries should include as many important facts as possible, be easy to understand, and avoid repetition. The outputs are evaluated against predefined ground truths to check their relevance, coverage, and alignment with gold-standard summaries. This task requires handling large-scale, noisy, and diverse data to produce useful and reliable summaries.

2. Dataset

The CrisisFACTS dataset serves as the cornerstone of this research, providing a detailed and structured framework for the end-to-end process of data retrieval, ranking, and summarization.This dataset is designed to capture the multifaceted nature of crises and ensure alignment with the diverse needs of stakeholders. It is organized into four key components, each serving a distinct purpose within the workflow.

1. **Event Definitions:** This component establishes the context for each crisis by categorizing and defining events. Key columns include **eventID** (a unique identifier),**title** (a concise name for the event), **type** (e.g., wildfire, earthquake), and **description** (a detailed summary of the event).

2. **User Profiles:** User profiles capture stakeholders' information needs, focusing on their specific queries. The maincolumns are **queryID** (a unique identifier for each query), **indicativeTerms** (key phrases definingrequirement).

3. **Content Streams:** This component aggregates real-time updates from diverse sources such as social media, news outlets, and other platforms. Columns include **streamID** (a unique identifier for each content entry), **text** (the content or message), **sourceType** (the platform or source of the data, e.g., Twitter,Reddit), and **unixTimestamp** (the time of the update in UNIX format).

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

4. **Summary Requests:** Summary requests provide the temporal scope for summarization tasks, defining the time frames and event-specific requirements. Columns here include **requestID** (a unique identifier for the request), **eventID** (linking the request to a specific event), **startUnixTimestamp**, and **endUnixTimestamp** (defining the time window for the summary).

3. Approach

The CrisisFACTS framework collects crisis-related data from various sources like Twitter, Reddit, Facebook, and online news, organized by day. It processes short pieces of information (called "stream-items") to create a concise timeline of important details, filtering and simplifying the content as needed. Then we try to rank key "facts" for a specific event and day by **BM25** relevance scores for each document calculated, based on their importance for the summary. Next we evaluates the highest-ranked facts from each generated summary with golden summaries, focusing on what's most relevant to the event.

To transform vast, multi-stream datasets into concise and relatable summaries,we approached a two-stage process: relevant fact retrieval and ranking with summarization.

The first stage,**Relevant Fact Retrieval**, employs **BM25**, a probabilistic retrieval model, implemented via the pyTerrier framework. This stage is involved with filtering the overwhelming volume of raw data to extract a subset of content that aligns closely with the informational needs according to that particular scenario need. BM25 evaluates term frequency, document uniqueness, and length normalization to assign relevance scores, significantly reducing dataset size while preserving critical information. By narrowing the dataset to its most closest components, Thus,this stage helps the groundwork for efficient downstream processing.

In the second stage, **Ranking and Summarization**, advanced semantic methods are utilized to ensure the extracted content aligns closely with requested queries. **BERT**, a deep learning model for natural language understanding, computes semantic similarity scores that go beyond simple term matching, capturing nuanced relationships between queries and retrieved texts. This semantic scoring informs the ranking process, enabling the system to prioritize content of the highest relevance.

Two summarization techniques are used to cater to different use cases. **Extractive summarization** directly selects and organizes top-ranked facts, ensuring factual accuracy and timeliness, making it well-suited for real-time operational demands. In contrast, **abstractive summarization** leverages **BART**, a transformer-based encoder-decoder model, to generate fluent, coherent narratives that synthesize information into a more natural language form. This dual-path approach balances the need for both precision and readability, providing stakeholders with summaries that are both actionable and contextually rich.

4. Evaluation Metrics

The effectiveness of the proposed approach is evaluated using both traditional and advanced metrics, ensuring a assessment of its performance. ROUGE metrics, which measure textual overlap, are used to assess content recall and structural similarity. Specifically, **ROUGE-1** evaluates unigram overlap, ROUGE-2 measures bigram coherence, and ROUGE-L captures the longest common subsequence, providing a multi-faceted view of alignment with reference summaries.

BERT-Score, on the other hand, provides a deeper semantic evaluation by comparing the contextual embeddings of tokens in the generated and reference summaries. It calculates **precision**,**recall**, and **F1 scores** to gauge alignment at a semantic level, making it particularly well-suited for abstractive

summarization. These metrics collectively offer insights into the relevance, coherence, and contextual fidelity of the generated outputs.

Bert	ICS	NIST	WIKI
Bert_score/Rouge	0.4431/0.0418	0.5566/0.1307	0.5273/0.0267
TREC Best	0.4591/0.0581	0.5642/0.1471	0.5646/0.0362

Table 1
Bert Results

Baseline	ICS	NIST	WIKI
Bert_score/Rouge	0.4403/0.0398	0.5422/0.1229	0.5106/0.0248
TREC Best	0.4591/0.0581	0.5642/0.1471	0.5646/0.0362

Table 2
Baseline Results

BART	ICS	NIST	WIKI
Bert_score/Rouge	0.4428/0.0515	0.5271/0.1146	0.5046/0.0213
TREC Best	0.4591/0.0581	0.5642/0.1471	0.5646/0.0362

Table 3
BART Results

5. Results and Observation

The extractive summarized results show better performance compared to the abstractive summarized results. This difference happens because abstractive summarization uses multiple models to create summaries, but during evaluation, the gold summaries (the best possible summaries) are very detailed and specific. This makes it challenging for the abstractive models to match the level of detail and precision in the gold standards.

6. Key Challenges and Learning

The slow data retrieval from the server limits the system to fetch only the top 50 documents, which might miss other important information. This results in missing key facts spread across documents. Additionally, we did not consider temporal ordering, meaning the facts were not arranged based on their timestamps, which further impacts the completeness and accuracy of the summaries. Another challenge was that the golden summaries for all events were not fully accessible to participants and were only available to those in the TREC event. It would have been better if they were open to all participants. Lastly, abstractive summarization models like BART create natural summaries by synthesizing information but struggle to match the detailed gold summaries, leading to a mismatch with evaluation criteria that prioritize precision over fluency.

Key learnings included the Better performance of extractive summarization in aligning with structured reference summaries, the critical role of semantic scoring in enhancing content relevance,the importance of modular dataset design in adapting to varying crisis scenarios. Further more implementing and trying other approach like Ensemble technique may improve better summarization .These insights provide a strong foundation for future enhancements to the existing approach.

@daiict.ac.in

References

[1] L3S at the TREC 2022 CrisisFACTS Track. *TREC Proceedings*. Available at: https://trec.nist.gov/pubs/trec31/papers/eXSum22.R.pdf

[2] Dataset Source. *CrisisFACTS*. Available at: https://crisisfacts.github.io/

[3] CrisisFACTS: Building and Evaluating Crisis Timelines. *ISCRAM 2023 Proceedings*. Available at: https://crisisfacts.github.io/assets/pdf/crisisfacts2022.iscram2023.pdf

@daiict.ac.in