HiRAG: A Historical Information-Driven Retrieval-Augmented Generation Framework for Background Summarization

Dong Zhou DONGZHOU@GDUFS.EDU.CN
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, 510006, Guangdong, China

Binli Zeng 1521399204@QQ.COM
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, 510006, Guangdong, China

Nankai Lin[✉] NEAKAIL@OUTLOOK.COM
School of Computer Science and Technology,
Guangdong University of Technology, Guangzhou, 510006, Guangdong, China

Yongmei Zhou YONGMEIZHOU@163.COM
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, 510006, Guangdong, China

Aimin Yang AMYANG18@163.COM
School of Computer Science and Intelligence Education,
Lingnan Normal University, Zhanjiang, 524000, Guangdong, China

Abstract

In an era overwhelmed by a deluge of global information, it is often challenging for people to grasp the relationships that an event develops over time. The background summarization (BS) task facilitates a profound understanding of the relationships between the current background of an event at any given time and its historical backgrounds. To enhance comprehension and help news readers and professionals to quickly understand the evolution of events, we introduce a **H**istorical **i**nformation-driven **R**etrieval-**A**ugmented **G**eneration framework **(HiRAG)**. This framework is designed to extract the most relevant information from historical backgrounds and supplement it to generate precise background summarization. HiRAG employs state-of-the-art retrieval-augmented generation technologies to produce relevant background summarization. We implement a multi-strategy similarity calculation and introduce a sliding window mechanism to optimize retrieval construction. Our framework has been rigorously tested through a series of experiments and extensive analyses of the latest datasets. The promising results affirm the effectiveness of our proposed HiRAG framework and its retrieval capabilities.

Keywords: Retrieval Augmented Generation; Background Summarization; Similarity; Sliding Window Mechanism

1. Instruction

With the rapid advancement of technology, the volume of information generated globally is expanding exponentially, presenting people with unprecedented amounts of data. Consequently, swiftly and effectively extracting pertinent information has become increasingly challenging. Research such as (Babar and Patil, 2015; El-Kassas et al., 2021; Yang et al., 2023) addresses this issue by identifying and condensing the most crucial parts of extensive texts into concise, coherent summarizations, thereby enabling readers to grasp the core content quickly. In scenarios involving events driven by multiple occurrences, capturing the evolving relationships within an event becomes essential. Traditional document summarization methods are inadequate for these situations. Unlike conventional summarization, timeline summarization (TS) emphasizes the progression and interconnection of events along a given timeline (Swan and Allan, 2000). At the core of the TS task is the creation of a series of concise summarizations for news articles published across different periods (Li and Li, 2013), facilitating a quicker and more comprehensive understanding of news evolution.

Most research on TS tasks focuses on enhancing the performance of timeline summarization (Nguyen et al., 2014) or improving the summarization of events from time-stamped news articles (Yu et al., 2021). However, these research often overlooks that it is difficult for readers who first read a news event to quickly understand the entire content of a news event and that readers must quickly grasp the complexity of a large amount of information about new events in a short period of time. To address this issue, Pratapa et al. (2023) proposed a novel task, known as the background summarization (BS) task. The summarization in the BS task is a crucial tool for readers to grasp the relation of the latest news. These summarizations offer plentiful historical information relevant to the current background, effectively capturing the relationships with its past backgrounds. As illustrated in Fig. 1, each background summarization on the timeline effectively articulates the connections between historical and current backgrounds, bridging past developments with the current background.

When dealing with the timeline of news events, a significant challenge arises due to the extensive duration and numerous historical backgrounds. Large language models have become essential tools in natural language processing, excelling in complex language understanding and generation tasks due to their extensive internal corpus knowledge (Chang et al., 2023; Naveed et al., 2023; Minaee et al., 2024). These models perform exceptionally well on various tasks, such as summarization generation. However, despite their powerful language processing capabilities, they still face limitations in acquiring and precisely manipulating knowledge. Using all historical backgrounds as input to train a large model not only significantly increases the training time but also risks the model learning excessive information that is not directly related to the current background. This can result in background summarization that is inconsistent with the current background. As illustrated in Fig. 2, existing methods often generate background summarization using nearby historical backgrounds (Pratapa et al., 2023). However, this approach has clear shortcomings, leading the model to learn irrelevant information and increasing redundancy.

To address this issue, this paper focuses on the retrieval method that automatically filters out the most relevant and representative historical information for the current background. We propose a **H**istorical **i**nformation-driven **R**etrieval-**A**ugmented **G**eneration framework

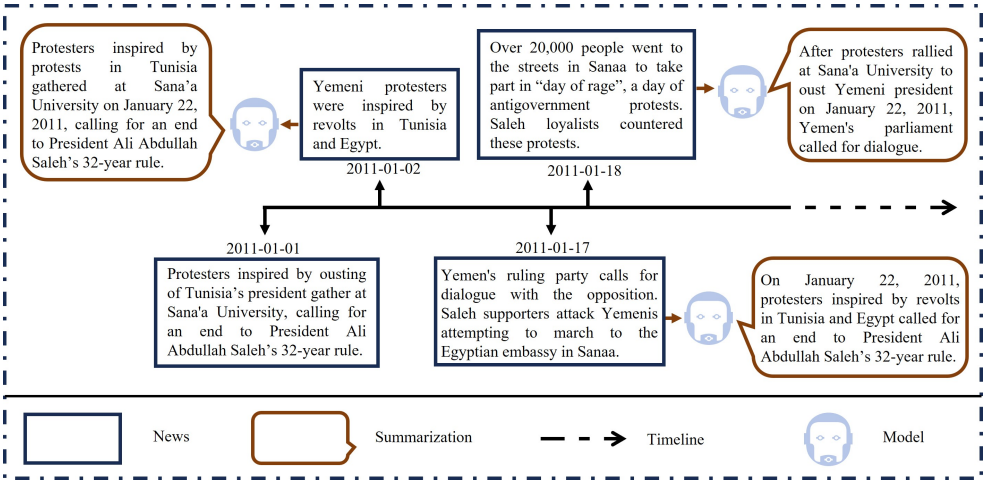

Figure 1: Illustration of the background summarization task.

(HiRAG), designed to enhance the generation of background summarization for event timelines. By leveraging retrieval-augmented generation (RAG) technology, the HiRAG framework effectively utilizes historical information to improve the quality of timeline background summarization. This approach not only helps readers gain a deeper understanding of the historical background of an event but also produces more precise and targeted background summarization by focusing on key information and current background.

In this work, our main contributions are as follows:

1. We propose the HiRAG framework, designed to generate background summarizations that are more consistent with factual accuracy.

2. We adopt a multi-strategy similarity calculation method and introduce a sliding window mechanism as a retriever. This enables the model to selectively extract the most pertinent backgrounds concerning recent events from extensive historical backgrounds, thus providing more accurate and detailed background summarization.

3. Experimental results from the LLAMA2-7B and LLAMA3-8B models confirm that HiRAG framework effectively guides models in producing background summarization that are more pertinent to the current background.

2. Relative Work

2.1. Timeline Summarization

Timeline summarization, a crucial tool for information retrieval and event organization, has consistently garnered significant attention. (Swan and Allan, 2000; Sipos et al., 2012) proposed an automated approach to process vast quantities of time-related textual data, extract event and time information, and organize it into a structured timeline summarization. Subsequent research focused more on generating timeline summarization, that is,

constructing or enhancing timeline (Allan et al., 2001; Althoff et al., 2015; Swan and Allan, 2000; Yan et al., 2011; Yu et al., 2021). Chieu and Lee (2004) first proposed that given a query, events related to the query are extracted from a collection of documents and placed in a timeline, and a key sentence is extracted from the collection to represent an event. Piskorski et al. (2020) employed an entity-centered event extraction method for extracting event timelines from online news. This method aims to clearly display the development of events in the form of a timeline and help users quickly understand the ins and outs of events.

2.2. Background Summarization

Dang et al. (2008) initially hypothesized that readers are familiar with the first ten articles, which enables the generation of new information about the topic for the following ten articles. In the work of Aslam et al. (2015), it was proposed that in some application scenarios, such as news events such as natural disasters or large-scale protests, which need to release backgrounds to users over time, an optimal background is to the minimum number of sentences covers all the information space of the event that the user may want to know. Wang et al. (2018) proposed a new news summarization method, which divides news articles into different event stages, such as event occurrence, event development, event results, etc., and then extracts the most critical information for each stage. Hayashi et al. (2020) proposed a decoupled paper summarization. One summarization specifically described the contribution of the paper, and the other summarization summarized the background and contextual information of the paper. This method allows readers to have a deeper understanding of the content of the paper. Pratapa et al. (2023) introduced a novel event timeline background summarization task, aimed at generating background summarization based on historical background information. This approach allows one to understand the relationship between the current background and its historical backgrounds by simply reading the current background summarization.

2.3. Retrieval-augmented Generation

Retrieval-augmented generation technology leverages the huge corpus in LLM to perform precise information retrieval, providing detailed background knowledge and rich contextual information for the text generation process. This technology enhances the reliability and accuracy of the text produced by large models, and it is extensively applied in open question and answer fields (Du and Ji, 2022; Siriwardhana et al., 2023; Hei et al., 2024), dialogue Thulke et al. (2021) and other aspects (Zeng et al., 2024; AI4Science and Quantum, 2023). Lewis et al. (2020) proposed a retrieval-augmented generation framework that integrated pre-trained language models with external knowledge bases, enabling the full utilization of relevant knowledge in the generation process. Parvez et al. (2021) proposed a framework for retrieval-augmented generation. A retrieval-augmented generation framework was proposed, which retrieved relevant codes or summarizations from the retrieval database and used them as a supplement to the code generation or summarization model to imitate the development at a time. Human code or summarization generation behavior makes it easier for developers to review it and improve efficiency. Wang et al. (2023) Interactively fused pre-trained large language models with external knowledge bases, and the self-knowledge-

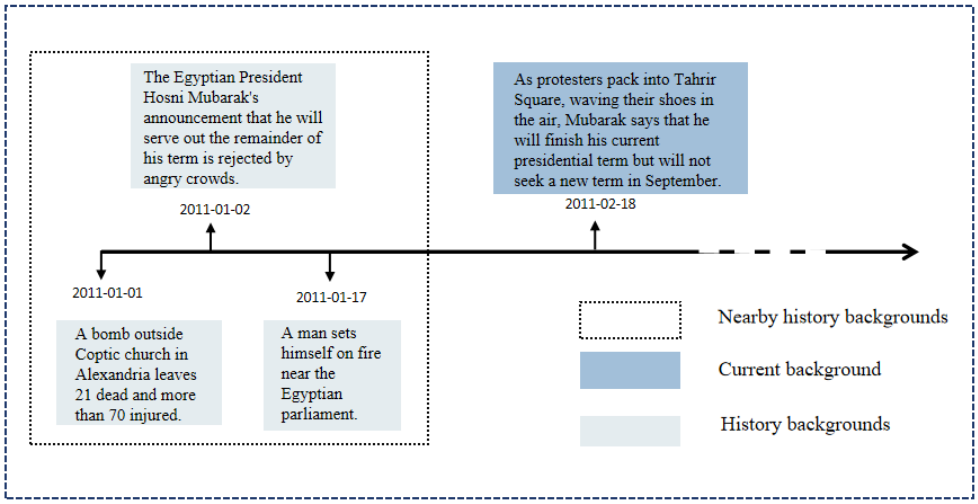

Figure 2: Illustration of the selection of the historical backgrounds adjacent to the current background.

guided fusion module can use the internal knowledge of the language model to effectively select and integrate retrieval results.

3. Method

As shown in Fig. 3., our proposed HiRAG framework fully utilizes the historical background information of events to generate comprehensive background summarization. By integrating historical background information, we enhance the coherence and accuracy of the generation of background summarization, allowing for a deeper understanding of the connections between the current background and its historical backgrounds.

3.1. Task Definition

Given an evolving series of background information, denoted as $<U_1, ..., U_T>$, our objective is to produce a sequence of background summarization, $<B_2, ..., B_T>$. Each summarization B_t corresponds to the accumulated historical backgrounds from U_1 through U_t. For each background U_t, our goal is to craft the most informative prompt template based on our custom-built retriever and generate a summarization, B_t, that is intimately aligned with the current background.

In executing this process, we utilize the current background, U_t, to guide the retriever toward extracting the most pertinent historical backgrounds. This strategy ensures that each generated summarization B_t is not only highly relevant to the current background but also accurately reflects the interplay between historical and current backgrounds, thus providing a deeper understanding of the evolving narrative.

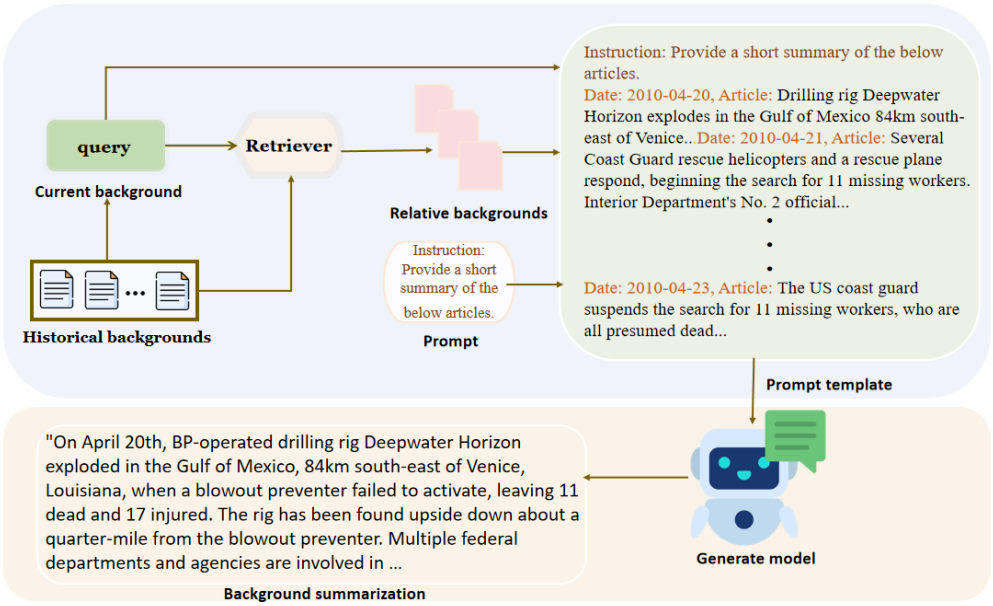

Figure 3: Illustration of the HiRAG framework.

3.2. Retriever

Our retriever incorporates three finely segmented modules. Initially, we employ a sliding window mechanism to define the scope for retrieving historical background information. Subsequently, we utilize a multi-strategy similarity calculation to assess and select the most relevant historical backgrounds based on their similarity scores. Finally, these chosen backgrounds are integrated with common task prefixes to construct a prompt template that guides the model in generating summarizations. An overview of these modules is presented in this section.

The complete operational workflow of our retriever is illustrated in Fig. 4. Specifically, we process the background through entity extraction and word segmentation. Subsequently, we apply a multi-strategy similarity calculation to the historical background information within the specified window. This process involves scoring each historical background similarity based on $U(t)$ and selecting the highest score as the subsequent query. The diagram suggests that the most relevant historical background is identified as $t^{'}$, prompting an adjustment of the window range to $[t^{'}-L, t^{'}+L-1]$. Through this cyclical process, we identify the TOP K most relevant historical backgrounds, which are then concatenated with the task prefix to create a comprehensive prompt template.

3.3. Historical Backgrounds Scoring

In geometric space, entities and texts are positioned within a multi-dimensional space where similarity measures the relationships between entities as well as between texts (Szmeja et al., 2018). Entity similarity typically quantifies the likeness between two or more entities, while text similarity is generally employed to compare two texts or multiple text segments. Our objective is to identify the most relevant and informationally complementary historical back-

ground to the current background. To achieve this, we employ a multi-strategy approach that finds historical background with high entity similarity but low text similarity.

Initially, by loading an English model through SpaCy, we utilize named entity recognition (NER) technology for entity extraction and tokenization of the backgrounds, and process backgrounds by tokenizing and subsequently removing stop words. Subsequently, we use the current background, U_t, as a query q, and apply the Jaccard similarity calculation to evaluate the relationships between entities and texts within their historical backgrounds relative to q. The formula for this calculation is:

$$es = \frac{\{e_q \mid e_q \subset \text{EntityExtraction}(q)\} \cap \{e_{U_i} \mid e_{U_i} \subset \text{EntityExtraction}(U_i)\}}{\{e_q \mid e_q \subset \text{EntityExtraction}(q)\} \cup \{e_{U_i} \mid e_{U_i} \subset \text{EntityExtraction}(U_i)\}} \tag{1}$$

$$ts = \frac{\{w_q \mid w_q \subset \text{Tokenizer}(q)\} \cap \{w_{U_i} \mid w_{U_i} \subset \text{Tokenizer}(U_i)\}}{\{w_q \mid w_q \subset \text{Tokenizer}(q)\} \cup \{w_{U_i} \mid w_{U_i} \subset \text{Tokenizer}(U_i)\}} \tag{2}$$

Among the terms defined, e_q represents the entity extracted from the current background, while w_q denotes the vocabulary processed from the current background. The variable i serves as the index for historical backgrounds. Additionally, e_{U_i} is the entity extracted by the i-th historical background, and w_{U_i} refers to the vocabulary processed after the i-th historical background.

$$score = es - ts \tag{3}$$

The score we ultimately calculate quantifies the correlation between the current background and its historical backgrounds.

3.4. Historical Backgrounds Selection

The application of the sliding window mechanism in natural language processing predominantly spans various domains such as text analysis (Yao et al., 2019), information extraction (Huang et al., 2015), and sentiment analysis (Zhang et al., 2018). This mechanism enables models to concentrate on fixed-length segments of text—known as windows and progressively shift across the text to encompass it entirely. The Swin Transformer further refines this approach by limiting self-attention calculations to smaller windows (Liu et al., 2021), significantly reducing computational demands.

When dealing with exceedingly long event timeline summarization, we cannot overlook the issue that arises as the timeline extends: the information and significance of earlier background elements diminish (Chen et al., 2015). Consequently, historical backgrounds that are more distant from the current background, U_t, increasingly reduce their influence and relevance to U_t. To address this challenge, we have adapted this concept to our approach to background summarization.

Specifically, we initiate with a window size of L, setting U_t as *query*, and defining the initial window range as $[t-L, t-1]$. A scoring module is then employed to identify the historical background most pertinent to U_t within this range. After each selection of the most relevant historical background ($U_t^{'}$), we adjust *query* to $U_t^{'}$ and modify the window size to span $[t^{'}-L, t^{'}+L-1]$. It is crucial to respect historical boundaries, which we maintain between $[0, t-1]$. If extending $t^{'}+L-1$ surpasses $t-1$, we truncate it at $t-1$. Similarly, if $t^{'}-L$ falls below zero, we adjust it to zero.

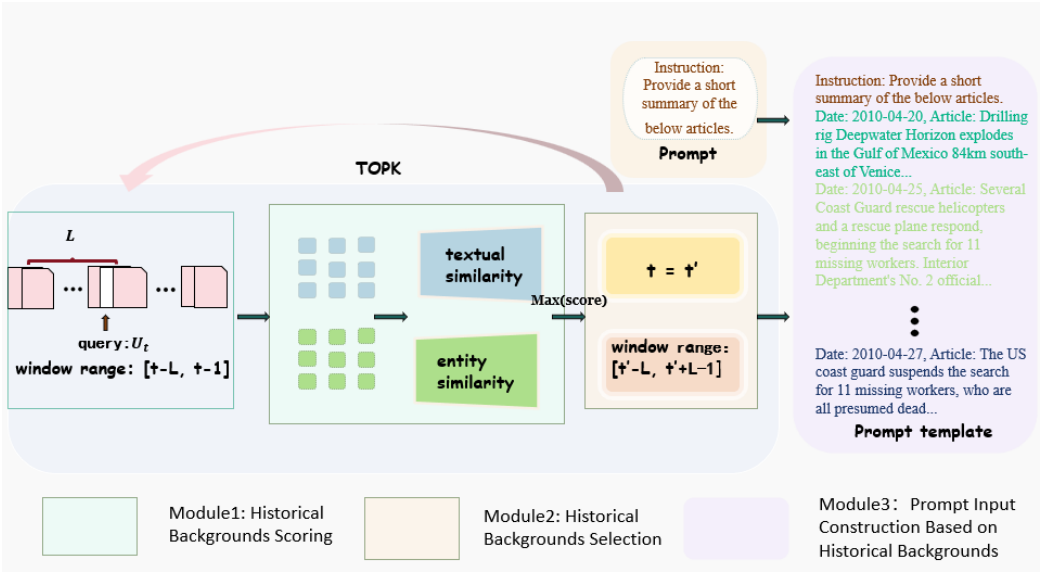

Figure 4: Illustration of the retriever. Here, L represents the initial window size, and t denotes the current background index, establishing the search range for historical backgrounds as $[t-L, t-1]$. In this setup, we use the t-th background as the query. The blue squares indicate the entity information extracted from the backgrounds, while the small green squares denote the word segmentation and processed segmentation data.

3.5. Prompt Input Construction Based on Historical Backgrounds

At a time when the issue of hallucinations in large models is becoming increasingly prevalent, the essence of RAG lies in the construction of effective prompt templates for these models. In this section, we employ the task prefix 'Provide a short summary of the below articles.' to enhance the accuracy of the summarizations produced. Furthermore, we concatenate this with relative historical backgrounds alongside the current background, U_t. Specifically, the prompt template is formulated as:'Instruction: Provide a short summary of the below articles. Input:<current background><relative historical backgrounds>'.

4. Experiments

4.1. Datasets

We utilize the dataset introduced by Pratapa et al. (2023) for background summarization, which has been meticulously annotated by experts across three well-regarded news timeline summarization datasets, encompassing 14 significant news events. The rigorous annotation process ensures that for each timestep in the timeline creation, there are three pairs of expertly rewritten backgrounds and corresponding background summarization.

Table 1: Effect of Retrieving Relative Backgrounds: STS and PSS are methods introduced in the latest paper, with the first four entries showing the results of their application on the Flan-T5-XL and GPT-3.5 models. Subsequently, we applied these methods along with HiRAG to the LLAMA2-7B and LLAMA3-8B models.

Method	ROUGE-1	ROUGE-2	ROUGE-L
Generic (Flan-T5-XL)	43.5	20.4	39.9
Generic (GPT-3.5)	40.5	15.5	36.6
Query-focused (Flan-T5-XL)	43.0	20.6	39.5
Query-focused (GPT-3.5)	40.2	15.4	36.1
STS (LLAMA2-7B)	21.8	3.9	19.9
STS (LLAMA3-8B)	23.7	4.7	21.2
PSS (LLAMA2-7B)	24.4	5.2	22.3
HiRAG-v1 (LLAMA2-7B)	29.2	7.3	26.7
HiRAG-v2 (LLAMA2-7B)	28.0	6.0	25.5
HiRAG-v1 (LLAMA3-8B)	30.6	8.5	27.3
HiRAG-v2 (LLAMA3-8B)	**32.3**	**9.1**	**28.9**

4.2. Experimental Settings

Generative Model For our task of background summarization, we employ two substantial open-source models: LLAMA2-7B and LLAMA3-8B as the backbone of the HiRAG framework.

- **LLAMA2-7B.** LLAMA2-7B represents a key installment in the LLAMA2 series of large language models (LLMs) developed by Meta AI, boasting a parameter scale of 7 billion (7B) Touvron et al. (2023). This model, both pre-trained and fine-tuned, demonstrates robust performance across a diverse array of NLP tasks.

- **LLAMA3-8B.** LLAMA3-8B represents a significant advancement over its predecessor, LLAMA2, in numerous respects. It has been trained on approximately 15 trillion tokens and supports text inputs up to 8,000 characters in length Meta (2024). This enhancement is particularly beneficial for processing extended text scenarios.

Evaluation Metrics In this paper, we employ the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) series of metrics to assess the quality of the summarizations generated by our system. The ROUGE index, as defined by Barbella and Tortora (2022), utilizes n-gram overlap to measure the similarity between the system-generated summarization and their corresponding reference summarization. We focus primarily on three specific metrics: ROUGE-1, ROUGE-2, and ROUGE-L, each of which evaluates summarization quality based on different n-gram lengths.

- **ROUGE-1** ROUGE-1 measures the extent of single-word (1-gram) overlap between the generated summarization and reference summarization. (Ganesan, 2018). Higher values in these metrics indicate greater lexical similarity between the generated summarization and the reference summarization.

- **ROUGE-2** ROUGE-2 evaluates the overlap of two consecutive words (2-grams) (Ng and Abrecht, 2015). This enables ROUGE-2 to capture phrase-level similarities that are not evident when only single-word comparisons.

- **ROUGE-L** ROUGE-L is based on the Longest Common Subsequence (LCS) approach (Schluter, 2017). This metric considers not only the overlap of words or characters but also their sequential order, thus placing a greater emphasis on the coherence and contextual consistency of the summarization.

Training Settings We propose guiding the construction of prompt templates under two settings, denoted **HiRAG-v1** and **HiRAG-v2**. The first, HiRAG-v1, involves arranging both the historical and current background in chronological order, followed by the addition of a task prefix to form the prompt template for the generation model. The second configuration, HiRAG-v2, directly inputs the retrieved historical information which directly creates a prompt template by appending a task prefix to the retrieved historical background before inputting it into the generation model. Both parameter exploration and ablation studies are conducted using the HiRAG-v1 method.

4.3. Comparison Method

We have compared several methods recently proposed in background summarization research. The specific methods include:

- **Generic**: Pratapa et al. (2023) use the task prefix 'summarize:' to guide the model in generating a summarization.

- **Query-focused**: Pratapa et al. (2023) employ 'Generate a short query-focused summary of the background.' as the task prefix and form the prompt with 'Query: <query>Background: <historical backgrounds>'. By combining the task prefix with the prompt, a prompt template is created to guide the model in generating summarization. The historical backgrounds selected are closely related to the current background, with the query representing the current background.

- **Single-Step Summarization (STS)**: According to the method proposed by Pratapa et al. (2023), we use the current background as the query and add a task prefix to create the prompt template, as follows: 'Instruction: Provide a short summary of the below articles. Input: <query>'.

Table 2: Exploration of the Number of Backgrounds Parameter.

Number of Backgrounds	ROUGE-1	ROUGE-2	ROUGE-L
3	24.7	5.5	22.7
4	26.5	6.2	24.2
5	27.5	5.7	24.9
6	**28.2**	**9.2**	**26.0**

- **Preceding-Steps Summarization (PSS)**: Following the method proposed by Pratapa et al. (2023), we choose the nearby historical backgrounds from previous backgrounds to incorporate into the prompt template, specifically: 'Instruction: Provide a short summary of the articles below. Input: <query><nearby historical backgrounds>'.

4.4. Main Results

We apply both comparative methods and our proposed methods on the two backbone models, LLAMA2-7B and LLAMA3-8B. According to the results shown in Table 1, the ROUGE scores of HiRAG-v1 and HiRAG-v2 are significantly higher than those of the baseline methods STS and PSS on both backbone models. These results demonstrate the effectiveness of our proposed HiRAG model in reducing redundancy in the selection of historical backgrounds and enhancing the model's focus on the information most relevant to the current background. On LLAMA3-8B, HiRAG-v2 outperforms HiRAG-v1, whereas, on LLAMA2-7B, the former is inferior to the latter. Although open-source language large models do not achieve the performance levels of current pre-trained language and proprietary large language models tailored for BS tasks, this nevertheless demonstrates that our method can improve the performance of these open-source models to a certain degree.

4.5. Parameter Exploration

Number of Backgrounds. The configuration of background quantity is intended to manage the complexity of the input document. To ensure optimal experimental outcomes with a reduced number of backgrounds, we conduct a series of experiments on the LLAMA2-7B model with background counts set at 3, 4, 5, and 6. The data shown in Table 2 demonstrate that selecting six backgrounds yields superior ROUGE scores compared to other configurations.

Window Size. The window mechanism proves to be an effective strategy for preventing the dilution of relevant and important information in the background as the timeline extends. Consequently, to optimize the window size, we conducted studies using three different window sizes—5, 10, and 15—on the LLAMA2-7B model. As indicated in Table 3, we observed that constraining the window size to 15 enhances the effectiveness of our method.

Table 3: Exploration of Window Size Parameters

Window Size	ROUGE-1	ROUGE-2	ROUGE-L
5	25.6	5.7	23.3
10	25.7	6.4	23.4
15	**26.5**	**6.2**	**24.2**

Table 4: Our comprehensive approach utilizes the HiRAG framework we proposed for training, where the 'w/o' sign denotes the omission of a specific feature within our framework. For instance, 'w/o Sliding Window Mechanism' indicates the removal of the sliding window mechanism from our HiRAG framework for experimental purposes.

Methods	ROUGE-1	ROUGE-2	ROUGE-L
HiRAG-v1	**29.2**	**7.3**	**26.7**
w/o Sliding Window Mechanism	26	6.2	23.8
w/o Entity Similarity	25.7	5.6	23.5
w/o Text Similarity	27.5	6.5	25.1

4.6. Ablation Study

As detailed in Table 4, we conduct experiments on the LLAMA2-7B model. Our findings reveal a considerable decrease in effectiveness when historical backgrounds are selected based solely on entity similarity or text similarity. Meanwhile, when using only a multi-strategy similarity approach, the ROUGE scores are substantially lower compared to results obtained from integrating the window mechanism. This underscores the efficacy of our proposed method.

4.7. Case Study

We conduct a case study to evaluate the strengths and weaknesses of our method in the BS task. The experimental results are presented in Fig. 5. HiRAG framework effectively generates background summarization by utilizing sufficient important information as support. However, it also tends to produce redundant information. In contrast, the comparison method PSS does not generate redundant information, but it has a tendency to directly copy content from the current background, focusing excessively on the present background.

5. Conclusion

We introduce the HiRAG framework, designed to maximize the use of historical background information for generating summarization highly relevant to the current background. We employ a sliding window mechanism to limit the retrieval range of historical backgrounds, effectively mitigating the reduced utility of distant historical backgrounds. Additionally, our method utilizes a multi-strategy similarity approach to identify the most pertinent historical background. Central to our approach is the deployment of a prompt template that melds a specific task prefix with dynamically retrieved content, including both current and historical backgrounds. This structured alignment of input data with the model's prompts sharpens the focus of the BS task, facilitating the creation of precise and appropriate background summarization. In future work, we will focus on optimizing our method to reduce redundancy in the generated summarization.

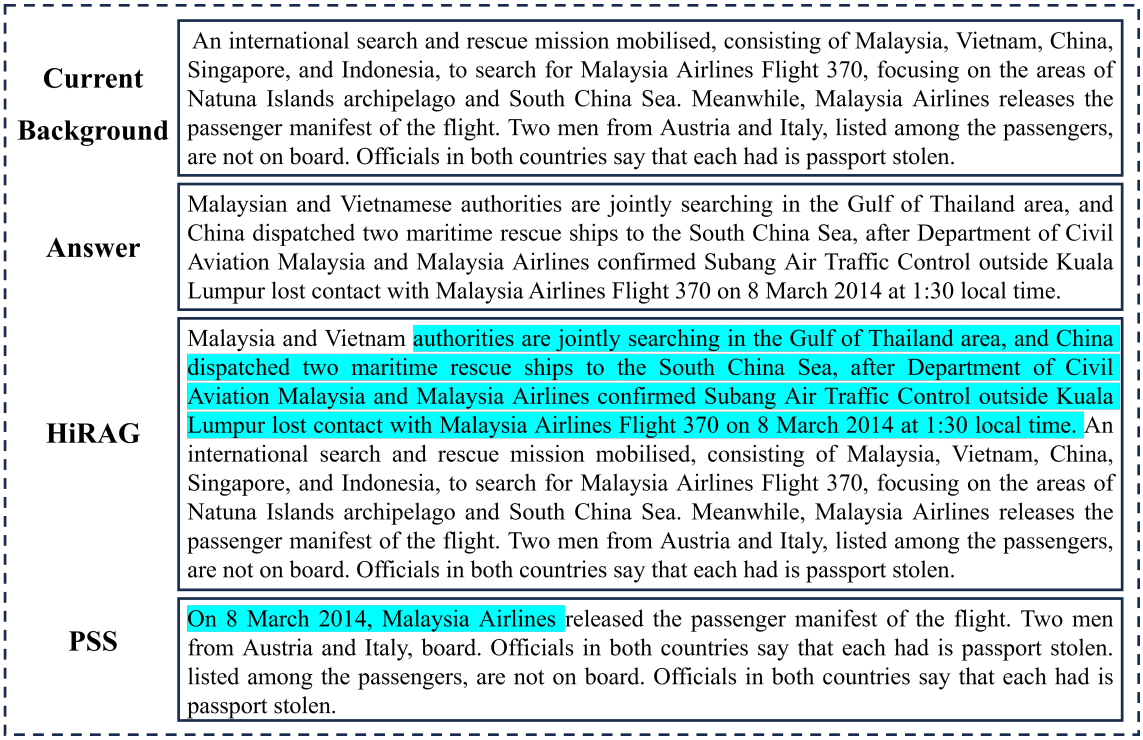

Figure 5: Prediction results of HiRAG and PSS for the example.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62376062), the Ministry of Education of Humanities and Social Science Project (No. 23YJAZH220, No. 24YJAZH244), the Philosophy and Social Sciences 14th Five-Year Plan Project of Guangdong Province (No. GD23CTS03), and the Guangdong Basic and Applied Basic Research Foundation of China (No. 2023A1515012718).

References

Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*, 2023.

James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18, 2001.

Tim Althoff, Xin Luna Dong, Kevin Murphy, Safa Alai, Van Dang, and Wei Zhang. Timemachine: Timeline generation for knowledge-base entities. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 19–28, 2015.

Javed A Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Richard McCreadie, Virgil Pavlu, and Tetsuya Sakai. Trec 2014 temporal summarization track overview. In *TREC*, 2015.

SA Babar and Pallavi D Patil. Improving performance of text summarization. *Procedia Computer Science*, 46:354–363, 2015.

Marcello Barbella and Genoveffa Tortora. Rouge metric evaluation for text summarization techniques. *Available at SSRN 4120317*, 2022.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Jie Chen, Zhendong Niu, and Hongping Fu. A multi-news timeline summarization algorithm based on aging theory. In *Web Technologies and Applications: 17th Asia-Pacific Web Conference, APWeb 2015, Guangzhou, China, September 18-20, 2015, Proceedings 17*, pages 449–460. Springer, 2015.

Hai Leong Chieu and Yoong Keok Lee. Query based event extraction along a timeline. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 425–432, 2004.

Hoa Trang Dang, Karolina Owczarzak, et al. Overview of the tac 2008 update summarization task. In *TAC*, 2008.

Xinya Du and Heng Ji. Retrieval-augmented generative question answering for event argument extraction. *arXiv preprint arXiv:2211.07067*, 2022.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165: 113679, 2021.

Kavita Ganesan. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*, 2018.

Hiroaki Hayashi, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. What's new? summarizing contributions in scientific literature. *arXiv preprint arXiv:2011.03161*, 2020.

Zijian Hei, Weiling Wei, Wenjie Ou, Juyi Qiao, Junming Jiao, Zhiqing Zhu, and Guowen Song. Dr-rag: Applying dynamic document relevance to retrieval-augmented generation for question-answering. *arXiv preprint arXiv:2406.07348*, 2024.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Jiwei Li and Sujian Li. Evolutionary hierarchical dirichlet process for timeline summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 556–560, 2013.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

AI Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*, 2024.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Jun-Ping Ng and Viktoria Abrecht. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*, 2015.

Kiem-Hieu Nguyen, Xavier Tannier, and Véronique Moriceau. Ranking multidocument event descriptions for building thematic timelines. In *COLING 2014, the 25th International Conference on Computational Linguistic*, pages 1208–1217, 2014.

Md Rizwan Parvez, Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. Retrieval augmented code generation and summarization. *arXiv preprint arXiv:2108.11601*, 2021.

Jakub Piskorski, Vanni Zavarella, Martin Atkinson, Marco Verile, et al. Timelines: Entity-centric event extraction from online news. In *Text2Story@ ECIR*, pages 105–114, 2020.

Adithya Pratapa, Kevin Small, and Markus Dreyer. Background summarization of event timelines. *arXiv preprint arXiv:2310.16197*, 2023.

Natalie Schluter. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 41–45. Association for Computational Linguistics, 2017.

Ruben Sipos, Adith Swaminathan, Pannaga Shivaswamy, and Thorsten Joachims. Temporal corpus summarization using submodular word coverage. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 754–763, 2012.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17, 2023.

Russell Swan and James Allan. Automatic generation of overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, 2000.

Paweł Szmeja, Maria Ganzha, Marcin Paprzycki, and Wiesław Pawłowski. Dimensions of semantic similarity. *Advances in Data Analysis with Computational Intelligence Methods: Dedicated to Professor Jacek Żurada*, pages 87–125, 2018.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *arXiv preprint arXiv:2102.04643*, 2021.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. Event phase oriented news summarization. *World Wide Web*, 21:1069–1092, 2018.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. *arXiv preprint arXiv:2310.05002*, 2023.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 433–443, 2011.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*, 2023.

Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.

Yi Yu, Adam Jatowt, Antoine Doucet, Kazunari Sugiyama, and Masatoshi Yoshikawa. Multi-timeline summarization (mtls): Improving timeline summarization by generating multiple summaries. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 377–387, 2021.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893*, 2024.

Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.