EventSum: A Large-Scale Event-Centric Summarization Dataset for Chinese Multi-News Documents

Mengna Zhu[1], **Kaisheng Zeng**[2,3], **Mao Wang**[1], **Kaiming Xiao**[1], **Lei Hou**[2*], **Hongbin Huang**[1*], **Juanzi Li**[2]

[1]Laboratory for Big Data and Decision, National University of Defense Technology
[2]Department of Computer Science and Technology, Tsinghua University
[3]College of Information and Communication, National University of Defense Technology
zhumengna16@nudt.edu.cn

Abstract

In real life, many dynamic events, such as major disasters and large-scale sports events, evolve continuously over time. Obtaining an overview of these events can help people quickly understand the situation and respond more effectively. This is challenging because the key information of the event is often scattered across multiple documents, involving complex event knowledge understanding and reasoning, which is under-explored in previous work. Therefore, we proposed the **E**vent-**C**entric Multi-Document **S**ummarization (**ECS**) task, which aims to generate concise and comprehensive summaries of a given event based on multiple related news documents. Based on this, we constructed the **EventSum** dataset, which was constructed using Baidu Baike entries and underwent extensive human annotation, to facilitate relevant research. It is the first large-scale Chinese multi-document summarization dataset, containing 5,100 events and a total of 57,984 news documents, with an average of 11.4 input news documents and 13,471 characters per event. To ensure data quality and mitigate potential data leakage, we adopted a multi-stage annotation approach for manually labeling the test set. Given the complexity of event-related information, existing metrics struggle to comprehensively assess the quality of generated summaries. We designed specific metrics including Event Recall, Argument Recall, Causal Recall, and Temporal Recall along with corresponding calculation methods for evaluation. We conducted comprehensive experiments on EventSum to evaluate the performance of advanced long-context Large Language Models (LLMs) on this task. Our experimental results indicate that: 1) The event-centric multi-document summarization task remains challenging for existing long-context LLMs; 2) The recall metrics we designed are crucial for evaluating the comprehensiveness of the summary information.

Code — https://github.com/Mzzzhu/EventSum
Extended version — https://arxiv.org/abs/2412.11814

Introduction

Dynamic events characterized by continuous development and change over time, uncertainty, and intricate causal relationships are pervasive in real life, such as natural disasters (*earthquakes, floods*), major sports events (*Olympics, UEFA European Championship*), and pressing social issues (*criminal cases, sudden public health crises*), etc. These events are often covered by multiple news articles that report from different perspectives and may include real-time updates. Integrating these diverse news sources is essential for a comprehensive understanding of the event. Extracting key information from related news articles to create accurate and comprehensive summaries is crucial for quickly organizing information about the event and better supporting downstream applications such as opinion mining, intelligent assistants, emergency response, etc (Tsirakis et al. 2017; Xu et al. 2020; Dutta et al. 2019; Purohit et al. 2018; Urologin 2018).

As illustrated in Figure 1, the news articles provide information about the "2023 Hebei Heavy Rain": the cause (News 1: Typhoon Doksuri and cold and warm air), casualties (News 3: 29 deaths), affected areas (News 4: 110 counties), and measures taken (News 5: all post-disaster reconstruction projects completed). According to the generated summary which combined the information from the above news articles, the reader can quickly and conveniently grasp the full scope of the "2023 Hebei Heavy Rain" event.

Generating concise and comprehensive summaries based on multiple documents surrounding the specified event not only makes the content more comprehensible to humans but also offers richer information. However, most of current research on event understanding is based on single-document and structural comprehension (Peng et al. 2023; Wang et al. 2022; Liu et al. 2020; Wang et al. 2020) and existing Multi-document Summarization (MDS) research faces three primary challenges: 1) News-focused datasets like Multi-News (Fabbri et al. 2019) consist of news-related articles and corresponding summaries that are organized around general news content rather than the specific dynamic event in chronological order; 2) Most large-scale datasets are automatically constructed, like WikiSum (Liu et al. 2018), which compromises dataset quality and increases the risk of data leakage in the test set; 3) Common evaluation metrics such as the ROUGE for lexical evaluation (Lin 2004) and BERTScore (Zhang et al. 2020) for semantic evaluation, are insufficient to adequately assess the completeness and comprehensiveness of summaries that focus on dynamic events.

To address the above challenges, in this paper, we constructed the first large-scale Chinese multi-news summariza-

*Corresponding Authors.

新闻 [1]	News [1]
……受台风“杜苏芮”和冷暖空气共同影响，自7月27日起，河北大部出现强降雨过程……	……Starting from July 27, heavy rainfall affected most of Hebei Province due to Typhoon Doksuri and the interaction of cold and warm air masses……
新闻 [2]	News [2]
截至8月1日12时……此次强降雨造成因灾死亡9人……因灾失踪6人……	As of 12:00 PM on August 1st…… the heavy rainfall has resulted in 9 deaths…… and 6 missing persons……
新闻 [3]	News [3]
……截至8月10日，河北因灾死亡29人，6人为先前失联人员……	……By August 10, 29 people had died in Hebei, including 6 previously missing persons……
新闻 [4]	News [4]
……8月11日，河北省政府新闻办召开“河北省防汛救灾暨灾后重建”新闻发布会……洪涝灾害波及110个县（市、区）……	……On August 11, the Hebei Provincial Government held a conference on flood control and disaster relief, reporting that the flooding affected 110 counties……
新闻 [5]	News [5]
……从省交通运输厅获悉，截至6月30日，我省交通领域2895项灾后重建工程全部完工……	……By June 30, all 2,895 post-disaster transportation reconstruction projects in Hebei were completed……
新闻 [N]	News [N]
……	……
摘要	**Summary**
2023年7月27日起，受冷暖空气和台风杜苏芮共同影响，河北省大部出现强降雨[1]……截至2023年8月10日，河北全省因灾死亡29人[3]……2023年8月11日，河北省政府新闻办召开“河北省防汛救灾暨灾后重建”新闻发布会通报，本次特大暴雨过程，洪涝灾害波及110个县（市、区）[4]……2024年7月，河北交通领域2895项灾后重建工程全部完工[5]……	Starting from July 27, 2023, heavy rainfall affected most of Hebei Province due to Typhoon Doksuri and cold and warm air masses [1]……By August 10, 29 people had died [3]……On August 11, it was reported that the flooding affected 110 counties [4]……By July 2024, all 2,895 post-disaster reconstruction projects in Hebei's transportation sector were completed [5]……

Figure 1: An example of the “2023 Hebei Heavy Rain”. The right side of the figure shows the English translation obtained from the original documents on the left.

tion dataset focused on dynamic events used for ECS, named EventSum. This dataset comprises a total of 5,100 events and 57,984 news articles, with each event corresponding to one piece of data. On average, each event has 11.4 related news articles and 13,471 input characters. In order to ensure the quality of the data, and the possibility of data leakage caused by the pre-trained corpus, we have implemented a multi-stage annotation method to manually write summaries for the test set data. These summaries are organized in chronological order and retain structured information obtained through annotation, including key sub-events, key event arguments, and causal relationships, which are crucial for comprehensive event understanding.

We used our dataset EventSum to evaluate the performance of several advanced long-text LLMs on this task. To better assess the quality of generated summaries and its effectiveness in organizing event information, we developed specific key elements recall metrics, including Event Recall, Argument Recall, Causal Recall, and Temporal Recall. We used existing structured event understanding datasets to train Natural Language Inference (NLI) models to compute these metrics by judging whether key event elements were entailed in the generated summary. This approach allows for a more comprehensive evaluation of the recall rate of structured information, as annotated in the dataset, thereby providing a more detailed measure of the quality of the generated summaries. The experimental results show that: 1) The event-centric multi-document summarization task remains challenging for current long-text LLMs on EventSum; 2) The designed recall metrics are significant for evaluating the comprehensiveness of the generated summaries.

Our contributions can be summarized as follows:

1. We proposed the event-centric multi-document summarization task which generated summaries around specified dynamic events based on given multiple related documents. This task is beneficial for quickly organizing event information but also challenging because it requires a deep understanding of long texts and complex event information.

2. We developed **EventSum**, the first large-scale Chinese multi-document summarization dataset, automatically constructed from Baidu Baike entries for this task study. Compared to existing news-related MDS datasets, EventSum features the highest number of input documents and the longest input length. To address data leakage and ensure robust evaluation, we **manually wrote** summaries for the test set and annotated key event-related information.

3. We conducted comprehensive experiments using annotated data of EventSum to evaluate the performance of advanced long-text LLMs on this task. Given the complexity of event data, we designed specific recall metrics to better assess the generated summaries. Our experimental results highlight the challenges posed by this task and dataset while confirming the effectiveness of our designed metrics.

Dataset Construction

In this section, we provide a detailed introduction to the construction methods of the EventSum. The overview of our method can be illustrated in Figure 2, which shows an example of the data construction process for the entry “2023 Hebei Heavy Rain”. It includes two parts: the automatic data construction process and the human annotation process.

Automatic Data Construction

The data for EventSum is sourced from event-related entries on Baidu Baike.[1] The description information from these entries are used as reference summaries. Due to the potential absence of references or the omission of critical references in these entries, the input sources for the summaries are derived from two components: 1) News articles that correspond to the references listed within the entry, and 2) News articles retrieved based on the title information of the entry. The primary sources of these input news articles are reputable official news websites such as CCTV News, Huanqiu, and Sina[2], which ensures the reliability of the information. The detailed construction process is as follows.

Data Collecting We initially employed web scraping tools such as Requests, BeautifulSoup, and Selenium to harvest

[1]https://baike.baidu.com

[2]https://www.cctv.com; https://www.huanqiu.com; https://www.sina.com.cn

entries related to notable events from Baidu Baike between January 2000 and April 2024. We collected and stored the "title", "card", "description", and "reference" for each entry. Non-event entries were filtered out based on key fields such as "time" and "location" in the basic information table, resulting in a total of 14,000 entries. To ensure the comprehensiveness of the summary information sources, we utilized the Bing News Search API[3] to retrieve related news articles for each entry based on its title. We specifically selected 20 news articles published within a month of the event date based on the "time" field in the card which contains basic event information as supplementary input documents.

Data Cleaning We cleaned and filtered the input documents through techniques like regular expression matching, removing some missing and duplicate documents. To further minimize noise introduced during document retrieval, we utilized the sentence-transformers[4] (Reimers and Gurevych 2020) to calculate the textual similarity between the retrieved documents and the summaries, filtering out low-relevance documents with a similarity score below the preset threshold of 0.5 and the number of input documents was controlled between 5 and 20.

Temporal Relation Annotation Considering that temporal relationships are relatively simpler compared to other types of event information, and in the summaries generated around dynamic events, the events we are concerned with that have temporal relationships usually appear with clear time information indicators or obvious conjunctions. Therefore, we used LLMs[5] to automatically annotate temporal relationships in the reference summaries, and only annotated event pairs with clear "before" and "after" relationships. Considering the transitivity of temporal information, we only label events that are directly adjacent in time. We obtained 14,932 temporal relationships in total.

Through the aforementioned steps, we automatically construct data pairs, which can be represented as $i = [D, r, T]$, where D represents the set of input documents, r represents the reference summary, and T represents the set of temporal relationships automatically annotated based on r.

Finally, we obtain 5,100 instances and split them into training, validation, and testing sets. In EventSum, each instance corresponds to a dynamic event.

Human Annotation

Considering the efficiency and cost of manual summarization, we chose to manually annotate the test set of EventSum due to the inherent data leakage issues associated with open-source data. To guarantee the data quality, we adopted a multi-stage annotation method, replacing the original reference summary r in the test set with the manually written summary r'. Detailed annotation process is as follows.

Sub-events and Arguments Annotation Annotate structural sub-events and their relevant argument information related to the core dynamic event in each input document. The sub-events were annotated as **sentences** containing key event information, and the arguments we focused on annotating included "time", "location", "person", and "organization". The definitions of the event and arguments mentioned above are widely adopted, similar to those in ACE 2005 (Walker et al. 2006). However, it is important to note that our work does not define any specific event schema.

Summary Writing Write a summary for each input document. During the summarization process, should consider the structured event information annotated in Step 1 and use expressions from the input documents as much as possible.

Global Information Annotation Deduplicate structured event information from the annotations of each document and organize the summaries chronologically to generate a global summary and compile structured event information.

Causal Relation Annotation Annotate causal relationships between sentences in the global annotated sub-events. Following MAVEN-ERE (Wang et al. 2022), the annotated causal relationships primarily include "*cause*" and "*precondition*", where "*cause*" indicates a sufficient condition, and "*precondition*" indicates a necessary condition.

Through this multi-stage annotation method, we ensure that the annotated summaries contain more comprehensive, complete, and accurate event-related information. The annotated data is iteratively checked and corrected to ensure high annotation quality. In the annotated data, there are 2,345 sub-events, 4,787 arguments, and 1,107 causal relationships.

Ultimately, each instance in the testing set is represented as $i = [D, r', T, G]$, where r' represents the newly manually written reference summary, and G represents the manually annotated global structured information.

Quality Control To ensure data quality, annotators were divided into three groups with cross-checks during annotation. Project managers conducted random checks and resolved conflicts, while acceptance reviewers verified manager-approved data, calculated pass rates, and provided revision requirements. This process continued until the data achieved a 90% pass rate. The pass rate is calculated as (number of data meeting the criteria / total data) * 100%. Criteria: 1) documents are relevant to the event, and 2) the summary covers key event elements and organized in chronological order. We sampled 50 instances from both automatically constructed and human-annotated data for review. The pass rate for automatic data was 81%, and for human-annotated data was 93%. For temporal relationships annotated by LLMs, we reviewed samples to ensure proper labeling of sub-events with clear time indicators or conjunctions. When the labeled main temporal relationships reach 80%, the data point is qualified. The qualified rate is 83%.

Data Analysis

The final constructed EventSum dataset contains 5,100 instances and 57,984 news articles. We choose to compare EventSum with Multi-News (Fabbri et al. 2019), DUC

[3]https://www.microsoft.com/en-us/bing/apis/bing-news-search-api

[4]The model we used is paraphrase-multilingual-mpnet-base-v2.

[5]LLMs used here and in Designed Recall Metrics are glm-4-9b.

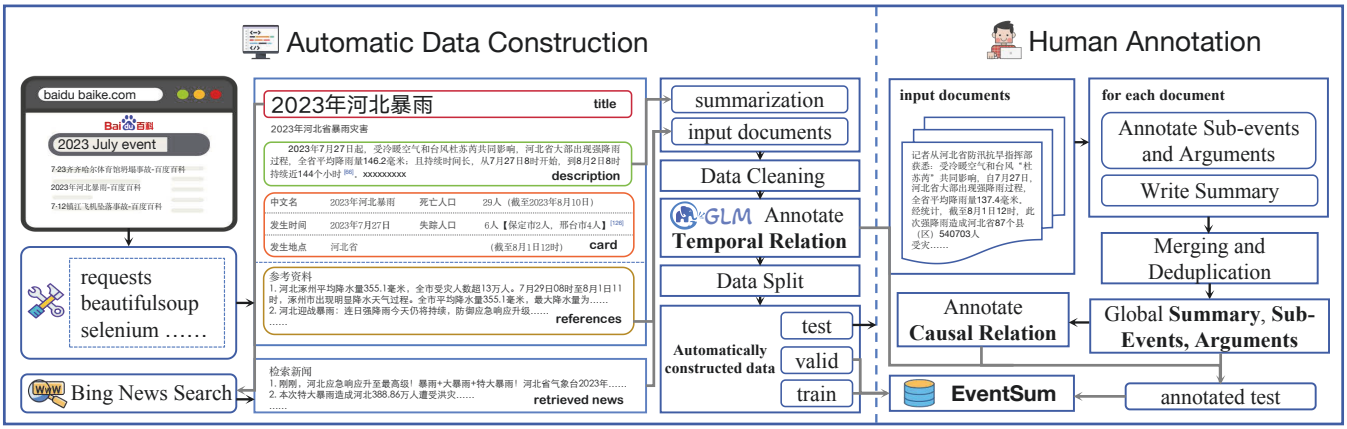

Figure 2: Overview of the construction process. It introduces the data construction process for the "2023 Hebei Heavy Rain" event from the retrieved entries for "Events of July 2023" on the Baidu Baike website.

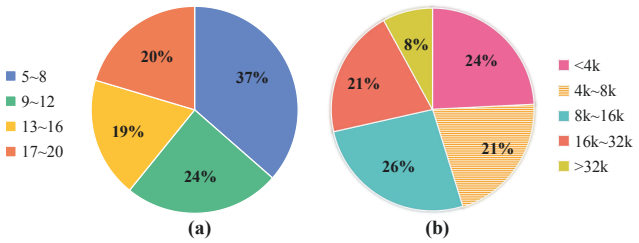

Figure 3: Analysis of Input Documents. The distribution of the number of input documents (a) and the total input characters (b) are presented on the left and right, respectively.

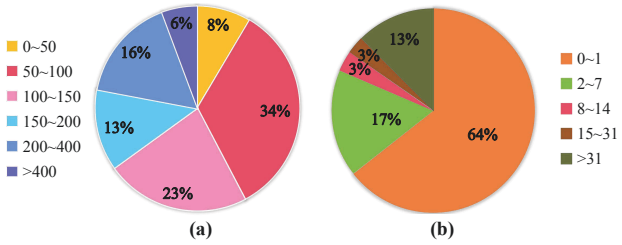

Figure 4: Analysis of Reference Summary. The distribution of characters in the reference summary is shown on the left (a), while the distribution of the dynamic event time span within the reference summary is displayed on the right (b).

data from 2003 and 2004 (Over and Yen 2004), and TAC 2011 (Karolina and Hoa 2011) data, which are typically used in multi-document settings and focused on news. The comparison result is shown in Table 1. It can be observed that EventSum is the first Chinese dataset specifically designed for multi-document summarization. Moreover, the number of input characters far exceeds that of other datasets. The number of input documents is also significantly higher compared to the widely used Multi-News.

To better understand the characteristics of EventSum, we conducted a detailed statistical analysis on the input news documents and the reference summary as follows.

Analysis of the Input Documents The number of input documents was controlled between 5 and 20. The average number of input documents is 11.4 and the distribution of the number of input documents can be seen in Figure 3 (a). Most instances have more than 8 input documents and one-fifth of the instances have more than 16 input documents.

The average input length is 13,471 characters, with a maximum length of 174,152 characters. The distribution of input lengths is shown in Figure 3 (b). Over half of the instances contain more than 8,000 characters in the input documents, and nearly one-third have more than 16,000 characters.

Analysis of the Reference Summary We conducted an analysis of the length distribution of the reference summaries, as illustrated in Figure 4 (a). The average length of the summaries is 161 characters, which meets the requirement for conciseness in practical applications. Additionally, we also analyzed the time span corresponding to the dynamic event in the reference summaries, as shown in Figure 4 (b). Nearly 40% of the data spans more than one day, and 13% spans more than one month, reflecting the distribution of events in real-world scenarios.

Evaluation Metrics

Evaluating metrics are essential for measuring the quality of the generated summaries in the summarization task, and well-defined metrics are crucial for relevant research (Ma et al. 2022). In this section, we introduce the evaluation metrics employed in our study in detail, including commonly used existing metrics and our specifically designed metrics.

Common Metrics ROUGE is the most commonly used metric in the summarization community and it comprises a set of evaluation metrics that assesses the similarity between the generated summary s and the reference summary r. It includes multiple variants to evaluate candidate summaries in different ways, with the most commonly used being ROUGE-N and ROUGE-L. ROUGE-N measures the n-gram recall between s and r. ROUGE-1 and ROUGE-2

Dataset	Language	Data Size	#Docs	#Words (Input)	#Words (Output)
DUC 03+04	English	320	10	4,636	110
TAC 2011	English	176	10	4,696	100
Multi-News	English	44,972/5,622/5,622	2.7	2,104	264
EventSum	Chinese	4,015/500/585	11.4	13,471	161

Table 1: Comparison between EventSum and existing MDS datasets with existing datasets that are focused on news and most similar to our data. Words represent tokens for English datasets and characters for Chinese datasets.

are special cases of ROUGE-N that are usually chosen as best practices and represent the unigram and bigram, respectively. ROUGE-L adopts the longest common subsequence algorithm to count the longest matching vocabularies.

BERTScore is a prominent semantic matching metric that leverages pre-trained BERT embeddings to compute the similarity between the tokens in the generated summary s and the reference summary r. This approach provides a context-aware measure of semantic equivalence.

In our paper, we used F1-scores of ROUGE-1, ROUGE-2, ROUGE-L and BERTScore for evaluation.

Designed Recall Metrics Given that our task focuses on event-centric summarization, in order to better evaluate the completeness and accuracy of event information in the generated summaries, we specifically designed key element recall metrics, including *Event Recall*, *Argument Recall*, *Causal Recall* and *Temporal Recall*.

Due to the diversity of text generation, it is not feasible to directly calculate the recall rate of key elements using simple methods like regular expression matching. Drawing inspiration from the Recognising Textual Entailment task which defines textual entailment as that one text fragment can be inferred from another text fragment (Dagan, Glickman, and Magnini 2005), we obtain relevant recall rate by judging whether key elements were entailed in the generated summary or not. The general formula for key elements is:

$$\text{Recall}_{k_i} = \frac{\sum_{e \in \mathcal{E}_{k_i}} \Gamma(e, s)}{|\mathcal{E}_{k_i}|}, \tag{1}$$

$$\Gamma(e, s) = \begin{cases} 1, & if\ e \subseteq s, \\ 0, & else. \end{cases} \tag{2}$$

Here, e represents the key element annotated based on the reference summary, $\mathcal{E}_{k_i}$ denotes the set of relevant annotated key elements with type k_i, and s denotes the generated summary. $\Gamma(e, s)$ is a discriminator used to determine whether the annotated element e can be inferred from the generated summary s or not. The entailment is confirmed only when $\Gamma(e, s) = 1$, indicating that the element e is indeed entailed in the summary s. The $\subseteq$ means the entailment relationship.

In the equation (1) and (2), the corresponding e to designed metrics *Event Recall*, *Argument Recall*, *Causal Recall* and *Temporal Recall* is the sentence containing key event information, event-related arguments, causal relationships, and temporal relationships respectively.

We used existing event understanding datasets and the automatically constructed data of EventSum to train a relevant binary classification Natural Language Inference (NLI) model as the discriminator. Specifically, CMNEE (Zhu et al. 2024) is used to construct data for training NLI models for Event Recall and Argument Recall as it is a large-scale Chinese event extraction dataset and annotates coref-arguments information such as abbreviations, pronouns, etc. Considering there is currently no suitable Chinese dataset for event causal relationships extraction study and causal relationships are relatively more complex, making it difficult to obtain reliable annotation automatically, we chose to use the translated version of MAVEN-ERE (Wang et al. 2022) to construct data for training the NLI model for Causal Recall because the causal relation annotation requirement is similar to that of EventSum. Additionally, the automatically constructed data of EventSum, where temporal relationships were annotated during the dataset construction process, was used to train the NLI model for Temporal Recall.

The structural annotation information is converted into natural language expression by LLMs as t_2 for positive instances, with input text designated as t_1 for the NLI models. To better evaluate the quality of the summaries, we analyzed the generated summaries and designed three strategies to construct negative instances by modifying a certain proportion of the positive instances, making the constructed data more closely aligned with our actual requirements. The negative instances generation strategies are as follows.

- **Remove**: Use the sentence-transformers library to evaluate the similarity between sentences in t_1 and t_2. Sentences from t_1 with a similarity score exceeding a threshold of 0.5 should be removed. The remaining sentences are then concatenated to form a new text, denoted as t_1'.
- **Revise**: Instruct the LLM to modify key event-related information in t_2, such as "time", "location", "quantity" and "person", etc. Alternatively, the LLM may expand upon or remove certain details surrounding the key event in t_2. The modified sentence is then used as t_2'.
- **Replace**: Randomly retrieve 100 instances, calculate the similarity between the text of the retrieved instances and t_2, and use the text with the highest similarity but no overlapping event information as the new text t_1'.

Then we trained models to determine whether t_2 (t_2') was entailed in t_1 (t_1') or not. The model we selected to train was chinese-roberta-wwm-ext (Cui et al. 2021) which provides robust support for the Chinese corpus.

On the test set of our constructed data for NLI models, final $Event\ Recall$, $Argument\ Recall$, $Causal\ Recall$, and $Temporal\ Recall$ are 96.8, 92.9, 94.3, 92.1 respectively.

We used trained models as discriminators to determine whether key elements annotated in the test set of EventSum were entailed in the generated summaries. The generated summary is as "t_1" and the key element is as "t_2". In this way, we could obtain all the recall metrics we need.

Experiments

In this section, we used the annotated data of EventSum to evaluate the performance of advanced long-context LLMs on the event-centric multi-documents summarization task.

Experimental Setup

We evaluate 10 popular LLMs that feature long context capability and good support for Chinese, including open-source models: Baichuan2-13b-chat (Baichuan 2023), Llama3-chinese-8B-Instruct (Cui, Yang, and Yao 2023), Yi-1.5-9b-chat-16k (Young et al. 2024), Qwen2-7b-Instruct (Yang et al. 2024), glm-4-9b-chat (GLM et al. 2024), InternLM2.5-7B-Chat-1M (Cai et al. 2024), glm-4-9b-chat-1M (GLM et al. 2024), and commercial models: MoonShot (Qin et al. 2024), Claude-3-Opus (Anthropic 2024), GPT-4o[6] (OpenAI 2024). Metrics we used have been introduced above. The assessment was conducted under the zero-shot setting.

Overall Results

Overall experimental results indicated that the task and our dataset are challenging, as shown in Tabel 2, We summarized our findings from following aspects:

1) Performance on Commonly Used Metrics: Among the open-source models, the best model was *glm-4-9b-1M*, while the best commercial model was *GPT-4o*. There is still significant room for improvement in overall performance. Open-source models outperformed commercial models on commonly used metrics. This is mainly because these models have undergone targeted training in Chinese, resulting in more natural and fluent expression in our task.

2) Effect of Input Length Limitation on Performance: Comparing the performance of open-source models with different input length limitations revealed that longer input length improved performance. This can be obviously seen from the results of *glm-4-9b-chat* and *glm-4-9b-chat-1M*.

3) Performance on Event-Centric Metrics: We observed a trend opposite to that seen with common metrics. *Claude-3-Opus* performed best on almost all our designed metrics, and commercial models generally performed better than most open-source models, which indicates the importance of our designed metrics for comprehensive evaluation.

4) Analysis of Our Designed Metrics: Specifically analyzing our designed metrics, we found that Event Recall was significantly lower compared to other metrics. This is mainly because the expression of sub-events is more complex than that of arguments and there is a greater quantity of sub-event data compared to relationships data.

[6]The models we used: moonshot-v1-128k, Claude-3-Opus-20240229, GPT-4o

Further Analysis

In this section, we randomly sampled 50 instances from the prediction results of the best-performing model, *Claude-3-Opus*, for manual observation to gain further insights into EventSum. Additionally, we analyzed the impact of the number of input documents and different time spans on performance, and assessed the reliability of the evaluation metrics to ensure a comprehensive analysis.

Analysis of Generated Summaries

To better understand the challenges of EventSum, we observed the generated summaries of the sampled data and summarized the common issues into 3 main categories:

1) Incomplete or Missing Information: The summaries might omit key elements of the dynamic event including sub-events, arguments, causal relations, and temporal relations mentioned above. This can lead to summaries lacking a comprehensive description of the dynamic event, as indicated by the results of our recall metrics.

2) Over or Under generalization: Summaries may be too vague, failing to capture specific details of the dynamic event, or too detailed, making the summary unnecessarily long. Striking the right balance between detail and brevity is a common challenge for the summarization task.

3) Irrelevance: Summaries might include irrelevant information that is not directly related to the dynamic event like reflective or commentary content or even other events information, especially when news like summary reports include multiple similar events appear in the input documents.

Additionally, in some models with poorer performance, issues such as repetition, incoherence, and poor responsiveness to the instructions in the prompt were also observed.

Metric Evaluation

Referring to the evaluation method of ROUGE, we compared the recall metrics obtained by trained NLI models with manually computed results based on human evaluation and also calculated their consistency on our sampled data to assess the reliability of our designed metrics, as shown in Table 3. The results in the table are close and relevant consistency all exceeds 90%. Most metrics from trained NLI models are generally slightly lower compared to human evaluation. This is because the high degree of diversity in the text generated by LLMs makes it difficult to identify some entailment relationships. However, this does not affect the comparison of the capability of LLMs, and the differences between the results are within an acceptable range. It indicates that our trained models can effectively compute the recall rate of key elements and better evaluate the quality and completeness of the summary.

Impact of Number of Input Documents

The impact of the Number of Input Documents can be seen in Figure 5 (a). It shows that almost all metrics exhibit clear overall downward trends. However, the trends for temporal and causal relationships recall differ from the other metrics. The best performance is not achieved with the fewest input documents. This suggests that the model may better capture

Type	Model	Length	R-1	R-2	R-L	BS	ER	AR	CR	TR
Open-Source	Baichuan2-13b-chat	8K	30.3	15.9	22.7	66.6	15.7	21.9	23.3	18.6
	Llama3-chinese-8B-Instruct	8K	40.1	22.8	29.8	73.5	15.1	26.9	24.4	20.2
	Yi-1.5-9b-chat-16k	16K	35.7	18.2	19.8	68.0	14.8	31.4	50.3	39.6
	Qwen2-7b-Instruct	32K	47.2	26.6	32.1	75.8	**27.4**	48.4	66.7	53.5
	glm-4-9b-chat	128K	48.2	27.0	35.6	**77.2**	16.2	32.8	22.8	15.3
	InterLM2.5-7B-Chat-1M	1M	47.9	26.7	33.7	76.3	24.5	44.2	55.2	40.5
	glm-4-9b-chat-1M	1M	**49.3**	**28.6**	**36.2**	77.0	23.9	43.8	47.5	31.8
Commericial	MoonShot	128K	43.2	23.2	29.9	71.9	23.5	43.7	57.9	42.3
	Claude-3-Opus	200K	45.1	22.9	29.7	75.2	25.7	**50.3**	**67.3**	**56.8**
	GPT-4o	128K	47.5	26.1	33.1	76.2	21.7	46.2	56.1	40.0

Table 2: Experimental results on EventSum. Length: Input length limitation of models; R-1: ROUGE-1; R-2: ROUGE-2; R-L: ROUGE-L; BS: BERTScore; ER: Event Recall; AR: Argument Recall; CR: Causal Recall; TR: Temporal Recall. Metric Definitions were illustrated in the Evaluation Metrics section. The best results are in bold. The second-best results are underlined.

Metric	Predicted	Human	Consistency
Event Recall	19.2	24.2	95.0
Argument Recall	36.4	39.4	97.0
Causal Recall	67.7	71.7	90.9
Temporal Recall	50.5	49.5	92.9

Table 3: Recall metrics obtained by trained NLI models and human judgment followed by consistency between them.

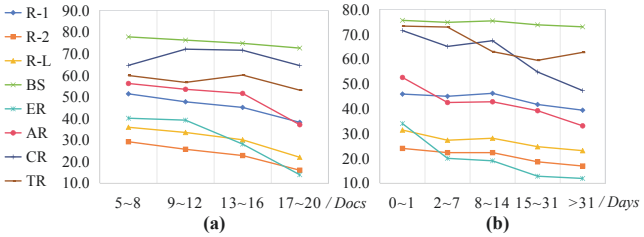

Figure 5: Analysis of the impact of various input documents number (left) and time span (right) of the dynamic event.

the relationships between events within a certain range of input document quantities, but as the number of input documents further increases, the complexity of the information rises, leading to a decline in the performances.

Impact of Time Span

The impact of time span of the dynamic event can be seen in Figure 5 (b). It can be observed that all metrics show an overall downward trend. The decline is more pronounced in our designed recall metrics compared to the common metrics. As the time span of the event increases, the number of corresponding sub-events, event arguments, causal relationships, and temporal relationships also increases, making the event information more complex and challenging to ensure the comprehensiveness and completeness of the generated summary. The change in the BERTScore, a metric for semantic matching, is relatively small, indicating that LLMs generally maintain good semantic relevance.

Related Work

Our work can be seen as an important extension of Multi-document Summarization (MDS), which focuses on generating concise summaries from multiple documents related to a specific topic. In this section, we will introduce some representative datasets and evaluation metrics for MDS.

Datasets MDS prioritizes the capture of key information across documents and emphasizes content coverage without being constrained by specific temporal or event-based structures. Except for Multi-News, DUC and TAC datasets we compared with EventSum above that focused on News, there are also many other datasets for MDS, such as WCEP (Ghalandari et al. 2020) constructed based on Wikipedia, Multi-XScience (Lu, Dong, and Charlin 2020) focused on scientific articles, GameWikiSum (Antognini and Faltings 2020) focused on game vedio, etc. In MDS, there is a specific task similar to ours, namely Timeline Summarization (TLS) (Chieu and Lee 2004), which requires generating daily summaries and arranging them in chronological order. The mostly used dataset for TLS is Timeline17 (Tran et al. 2013) and Crisis (Tran, Alrifai, and Herder 2015). Timeline17 contains 17 topics and 19 timelines in total. Crisis has 5 topics and 22 timelines annotated in total. Available data for TLS is limited, which impedes relevant research. Our research can offer insights into how to construct TLS datasets and may even serve as a potential resource for TLS.

Considering ECS not only requires an understanding of temporal progression, like TLS, but also demands the ability to delineate the core event and sub-events, capture event relationships (co-reference, causal, etc.), resolve conflicting information, and integrate updates from multiple sources, which is helpful for obtaining a deep understanding of event dynamics while providing a clear and comprehensive view, existing MDS datasets are not well-suited for our task.

Evaluation Metrics Conventional evaluation metrics used in the MDS research can be mainly divided into two categories: 1) lexical matching metrics which evaluate the sim-

ilarity between generated summaries and reference summaries based on exact word overlaps, such as ROUGE (Lin 2004), BLEU (Papineni et al. 2002), Pyramid (Nenkova, Passonneau, and McKeown 2007); 2)semantic matching metrics which evaluate the meaning and contextual relevance of generated summaries beyond surface-level word overlaps, such as BERTScore (Zhang et al. 2020), Moverscore (Zhao et al. 2019), METEOR (Banerjee and Lavie 2005). These metrics have been shown to have a relatively low correlation with human judgments, especially for tasks with creativity and diversity (Zhang, Yu, and Zhang 2024). There are also some specialized metrics commonly used in TLS tasks. Except for the commonly used Rouge series scores including concat F1, Agree F1 and Align F1, Date F1 is mostly used for key date selection evaluation (Li et al. 2021; Hu, Moon, and Ng 2024). With the development of LLMs and their outstanding performances in various natural language processing tasks, a series of recent work has tried to use LLMs for evaluation (Wu et al. 2023; Liu et al. 2023, 2024).

To conclude, the existing evaluation metrics in MDS are not suitable for accurately evaluating the event content and cannot effectively assess the comprehensiveness of event information in the generated summaries.

Conclusions and Future Work

We proposed the first large-scale, event-centric summarization dataset for Chinese multi-news documents, EventSum, which was automatically constructed based on Baidu Baike entries, along with a manually annotated test set to mitigate the impact of inherent data leakage. Given the event-centric nature of EventSum, we designed recall metrics, including Event Recall, Argument Recall, Causal Recall, and Temporal Recall, to complement commonly used metrics in summarization tasks for evaluation. Experimental results demonstrated that this task and dataset are challenging, and further analysis confirmed the effectiveness and importance of our designed metrics. In the future, we plan to extend our approach to English corpora and increase the proportion of long time span events. Additionally, we hope to explore more sophisticated methods to conduct extensive experiments and further enhance performance.

Ethical Statement

This paper presents a new dataset, and we discuss some related ethical considerations here: (1) **Copyright Statement.** All data utilized in this work are publicly available and freely accessible, with no inclusion of proprietary or restricted data. The use of these datasets strictly adheres to the terms and conditions of their respective platforms. As such, this work does not involve any copyright infringement or related issues. (2) **Worker Treatments.** We hire annotators from a professional data annotation company, ensuring fair compensation with agreed-upon salaries and workloads. All employment terms are contractually bound and adhere to local regulations. (3) **Risk Control.** Given that the texts in our dataset EventSum do not contain private information and are sourced from open data, we believe EventSum poses no additional risks. To verify this, we manually reviewed a random sample of the data and found no concerning issues.

Acknowledgments

We thank all the anonymous reviewers and meta reviewers for their valuable comments, as well as all of our team members for their support and assistance. This work is supported by Beijing Natural Science Foundation (L243006) and Natural Science Foundation of China (62476150).

References

Anthropic, A. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. *Claude-3 Model Card*.

Antognini, D.; and Faltings, B. 2020. GameWikiSum: a novel large multi-document summarization dataset. *arXiv preprint arXiv:2002.06851*.

Baichuan. 2023. Baichuan 2: Open Large-scale Language Models. *arXiv preprint arXiv:2309.10305*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; et al. 2024. InternLM2 Technical Report. arXiv:2403.17297.

Chieu, H. L.; and Lee, Y. K. 2004. Query based event extraction along a timeline. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 425–432.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; and Yang, Z. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3504–3514.

Cui, Y.; Yang, Z.; and Yao, X. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*.

Dagan, I.; Glickman, O.; and Magnini, B. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, 177–190. Springer.

Dutta, S.; Chandra, V.; Mehra, K.; Ghatak, S.; Das, A. K.; and Ghosh, S. 2019. Summarizing microblogs during emergency events: A comparison of extractive summarization algorithms. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2*, 859–872. Springer.

Fabbri, A. R.; Li, I.; She, T.; Li, S.; and Radev, D. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1074–1084.

Ghalandari, D. G.; Hokamp, C.; Glover, J.; Ifrim, G.; et al. 2020. A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1302–1308.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Hu, Q.; Moon, G.; and Ng, H. T. 2024. From moments to milestones: Incremental timeline summarization leveraging large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7232–7246.

Karolina, O.; and Hoa, T. D. 2011. Overview of TAC 2011 Summarization Track. *National Institute of Standards and Technology*.

Li, M.; Ma, T.; Yu, M.; Wu, L.; Gao, T.; Ji, H.; and McKeown, K. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6443–6456.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, J.; Chen, Y.; Liu, K.; Bi, W.; and Liu, X. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 1641–1651.

Liu, P. J.; Saleh, M.; Pot, E.; Goodrich, B.; Sepassi, R.; Kaiser, L.; and Shazeer, N. 2018. Generating Wikipedia by Summarizing Long Sequences. In *International Conference on Learning Representations*.

Liu, Y.; Fabbri, A. R.; Chen, J.; Zhao, Y.; Han, S.; Joty, S.; Liu, P.; Radev, D.; Wu, C.-S.; and Cohan, A. 2024. Benchmarking Generation and Evaluation Capabilities of Large Language Models for Instruction Controllable Summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 4481–4501.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522.

Lu, Y.; Dong, Y.; and Charlin, L. 2020. Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8068–8074.

Ma, C.; Zhang, W. E.; Guo, M.; Wang, H.; and Sheng, Q. Z. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5): 1–37.

Nenkova, A.; Passonneau, R.; and McKeown, K. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2): 4–es.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Over, P.; and Yen, J. 2004. An introduction to DUC-2004. *National Institute of Standards and Technology*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Peng, H.; Wang, X.; Yao, F.; Wang, Z.; Zhu, C.; Zeng, K.; Hou, L.; and Li, J. 2023. OmniEvent: A Comprehensive, Fair, and Easy-to-Use Toolkit for Event Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 508–517.

Purohit, H.; Nannapaneni, S.; Dubey, A.; Karuna, P.; and Biswas, G. 2018. Structured summarization of social web for smart emergency services by uncertain concept graph. In *2018 IEEE International Science of Smart City Operations and Platforms Engineering in Partnership with Global City Teams Challenge (SCOPE-GCTC)*, 30–35. IEEE.

Qin, R.; Li, Z.; He, W.; Zhang, M.; Wu, Y.; Zheng, W.; and Xu, X. 2024. Mooncake: A KVCache-centric Disaggregated Architecture for LLM Serving. arXiv:2407.00079.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4512–4525.

Tran, G.; Alrifai, M.; and Herder, E. 2015. Timeline summarization from relevant headlines. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings 37*, 245–256. Springer.

Tran, G. B.; Tran, T.; Tran, N.-K.; Alrifai, M.; and Kanhabua, N. 2013. Leveraging Learning To Rank in an Optimization Framework for Timeline Summarization. In *SIGIR 2013 Workshop on Time-aware Information Access (TAIA'2013)*.

Tsirakis, N.; Poulopoulos, V.; Tsantilas, P.; and Varlamis, I. 2017. Large scale opinion mining for social, news and blog data. *Journal of Systems and Software*, 127: 237–248.

Urologin, S. 2018. Sentiment analysis, visualization and classification of summarized news articles: a novel approach. *International Journal of Advanced Computer Science and Applications*, 9(8).

Walker, C.; Strassel, S.; Medero, J.; and Maeda, K. 2006. ACE 2005 Multilingual Training Corpus.

Wang, X.; Chen, Y.; Ding, N.; Peng, H.; Wang, Z.; Lin, Y.; Han, X.; Hou, L.; Li, J.; Liu, Z.; et al. 2022. MAVEN-ERE: A Unified Large-scale Dataset for Event Coreference, Temporal, Causal, and Subevent Relation Extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 926–941.

Wang, X.; Wang, Z.; Han, X.; Jiang, W.; Han, R.; Liu, Z.; Li, J.; Li, P.; Lin, Y.; and Zhou, J. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1652–1671.

Wu, N.; Gong, M.; Shou, L.; Liang, S.; and Jiang, D. 2023. Large language models are diverse role-players for summarization evaluation. In *CCF International Conference*

on Natural Language Processing and Chinese Computing, 695–707. Springer.

Xu, R.; Cao, J.; Wang, M.; Chen, J.; Zhou, H.; Zeng, Y.; Wang, Y.; Chen, L.; Yin, X.; Zhang, X.; et al. 2020. Xiaomingbot: A Multilingual Robot News Reporter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 1–8.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; et al. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zhang, H.; Yu, P. S.; and Zhang, J. 2024. A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. *arXiv preprint arXiv:2406.11289*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhao, W.; Peyrard, M.; Liu, F.; Gao, Y.; Meyer, C. M.; and Eger, S. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 563–578.

Zhu, M.; Xu, Z.; Zeng, K.; Xiao, K.; Wang, M.; Ke, W.; and Huang, H. 2024. CMNEE: A Large-Scale Document-Level Event Extraction Dataset Based on Open-Source Chinese Military News. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3367–3379.