Enhancing Event-centric News Cluster Summarization via Data Sharpening and Localization Insights

Longyin Zhang, Bowei Zou, and Ai Ti Aw
Institute for Infocomm Research, A*STAR, Singapore
{zhang_longyin,zou_bowei,aaiti}@i2r.a-star.edu.sg

Abstract

This paper tackles the challenges of clustering news articles by main events (MEs) and summarizing these clusters, focusing on diverse languages and localized contexts. Our approach consists of four key contributions. First, we investigate the role of dynamic clustering and the integration of various ME references, including event attributions extracted by language models (LMs), in enhancing event-centric clustering. Second, we propose a data-sharpening framework that optimizes the balance between information volume and entropy in input texts, thereby optimizing generated summaries on multiple indicators. Third, we fine-tune LMs with local news articles for cross-lingual temporal question-answering and text summarization, achieving notable improvements in capturing localized contexts. Lastly, we present the first cross-lingual dataset and comprehensive evaluation metrics tailored for the event-centric news cluster summarization pipeline. Our findings enhance the understanding of news summarization across N-gram, event-level coverage, and faithfulness, providing new insights into leveraging LMs for large-scale cross-lingual and localized news analysis.

1 Introduction

With the explosive development of language models (LMs) (Chung et al., 2022; Touvron et al., 2023; DeepSeek et al., 2024; Dubey et al., 2024), some existing works argue that research on traditional text summarization is approaching or exceeding human excellence (Adams et al., 2023; Pu et al., 2023; Van Veen et al., 2024). Recently, more and more researchers turn to more challenging attempts like multi-document diversity summarization (Huang et al., 2024), hybrid multi- and cross-lingual summarization (Wang et al., 2023), unified **M**ulti-lingual, **C**ross-lingual and **M**ulti-documents **S**ummarization (MCMS) (Ye et al., 2024), and so forth. Notably, the trend of focuses shifting from compression ratio and coherence to summarization faithfulness and coverage is becoming increasingly obvious in the text summarization realm (Huang et al., 2024; Elhady et al., 2024; Peper et al., 2024; Olabisi and Agrawal, 2024a). Following this trend, we aim to improve MCMS on multiple performance indicators and bridge the gap between MCMS and its real-scenario application.

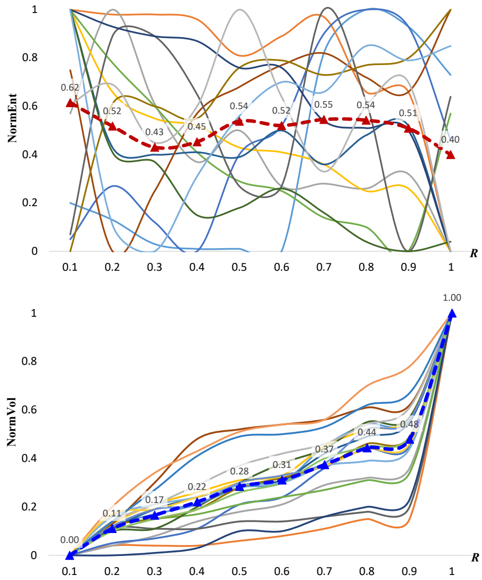


Figure 1: Distribution of information entropy and volume at different sentence sampling ratios (R). Solid lines represent randomly sampled article clusters, while the dotted lines represent the average results. The sentence sampling strategy is detailed in Subsection 3.2. For clarity, the entropy and volume values are normalized between 0 and 1; more details are shown in Appendix A.

News articles often cover interconnected events, blurring the boundaries between articles within a cluster and introducing uncertainty into downstream MCMS. To mitigate this error propagation, we propose a two-stage framework: event-centric news clustering followed by summarization. Specifically, we cluster articles based on their unique main event to establish clearer bound-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16412–16426
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

aries among clusters and then perform MCMS for each cluster independently. For MCMS, prior research has shown that large language models used for text summarization exhibit biases towards sentences with majority consensus (Lei et al., 2024), as well as the position bias (Olabisi and Agrawal, 2024b), hindering summarization diversity. Moreover, our analysis of several multilingual news clusters reveals an inverse trend between information entropy (Shannon, 1948) and volume (Huffman, 1952), shown in Figure 1. As we can see, the entropy value drops sharply at first, then continues to decline. Information volume, however, exhibits an accelerating increase with the sampling ratio. This suggests a potential correlation between the input text's information volume and entropy, and the MCMS results. We hypothesize that leveraging such correlation can improve the summarization performance, especially in this LLM era.

In summary, our contributions are three-fold:

- We introduce CLUST-MCMS, a novel two-stage pipeline for event-centric news clustering (ENC) followed by news cluster summarization, designed for real-world scenarios. Moreover, we contribute the first manually annotated evaluation data in this area, encompassing English, Chinese, Malay, and Indonesian.
- We enhance CLUST-MCMS with three key innovations: (1) a comprehensive investigation into finding the best main-event references for news clustering, (2) a data sharpening approach based on a novel analysis of the relationship between information volume, entropy, and summarization performance, and (3) an LM localization method to ensure summaries' faithfulness through temporal question-answering (TQA).
- This research aims for more than just brevity and fluency in the generated summaries. We manually annotated the key events of each article cluster for event-level coverage and entity-level faithfulness evaluation. Furthermore, to ensure a comprehensive evaluation of the pipeline task, we propose a robust metric that considers both the hit rate (F1) and the N-gram characteristics (ROUGE) of generated summaries.

2 Related Work

MCMS. Multi-document summarization (MDS) traditionally uses either abstractive (Zhang et al., 2018; Lebanoff et al., 2018; Gehrmann et al., 2018) or extractive (Angelidis and Lapata, 2018; Zheng et al., 2019; Mao et al., 2020) approaches to condense input news articles into a summary. In the past few years, the focus of MDS has shifted from primarily ensuring coherence and brevity (Peyrard et al., 2017; Chen et al., 2021) to prioritizing coverage and diversity (Amar et al., 2023; Ye et al., 2024; Huang et al., 2024), reflecting real-world needs such as social media news summarization. Furthermore, with increasing global interconnectedness, MDS is evolving from monolingual to multilingual and cross-lingual scenarios. Multilingual summarization generates summaries in the respective languages of the source articles (Giannakopoulos et al., 2015; Varab and Schluter, 2021; Feng et al., 2022), whereas cross-lingual summarization translates source articles into a target language for summarization (Yao et al., 2015; Ouyang et al., 2019; Wang et al., 2022). Recently, Ye et al. (2024) introduced a unified multi- and cross-lingual multi-document summarization task (MCMS), with multiple articles revolving around the same event. Following their work, we further extend the research to an event-centric news clustering and MCMS pipeline task (CLUST-MCMS) to address the dispersed nature of multilingual news online.

ENC. Event-centric news clustering (aka event detection) aims to group news articles discussing the same main event (Allan et al., 1998; Li et al., 2022; Liu et al., 2023). Due to the dynamic nature of social media news and limitations in data annotation, prior work often uses unsupervised clustering methods (Goyal et al., 2019; Carta et al., 2021; Liu et al., 2023; Lin et al., 2024). Leveraging LMs for event-centric news clustering remains relatively unexplored. (Tarekegn et al., 2024) represents an initial attempt in this direction, exploring LLM-driven keyword generation for improved clustering. Nevertheless, applying these methods to real-world scenarios, with its abundance of multilingual and cross-lingual content, presents significant challenges. While some research exists on cross-lingual news clustering (Miranda et al., 2018; Linger and Hajaiej, 2020; Santos et al., 2022; Wu et al., 2023), most work remains in finding better cross-lingual embeddings rather than directly addressing the inherent obstacles of cross-lingual clustering analysis. Our work addresses this gap by focusing on cross-lingual event-centric news clustering, specifically using LLMs to derive richer article representations beyond keywords for identifying main events.

3 Approach

3.1 ENC

Traditional text clustering methods usually group articles based on thematic similarity, typically using document vector representations. This work, however, focuses on event-centric news clustering, specifically using each article's main event (ME) as the clustering criterion. We define ME as the event sentence capturing the most significant action, occurrence, or issue within the news article. Besides, titles, lede sentences, or others conveying the most critical information are also prioritized as potential ME references.

ME generation. Leveraging the advanced natural language processing capabilities of recent LMs and addressing the growing need for open-domain news clustering solutions, we achieve ME generation using more accessible, smaller LMs. To ensure a high-quality and sustainable system, we utilize instruction-tuning data generated by GPT4 to fine-tune the smaller LMs for ME generation. Specifically, each ME is represented as a series of textual blocks representing varied event attributes:

$$S_e, S_t, S_{w*}, S_o = \mathrm{F}_{gpt}(S|A, I) \tag{1}$$

where S_e is a brief main event description, S_t the trigger, S_{w*} the combined WHO, WHERE, and WHEN information, S_o the event outcome, A the input news article, and I the prompt. We then fine-tune a small LM with the above data to generate a synthetic $\tilde{\mathrm{ME}}$ for each article. Besides, the title and lede sentences of an article also provide valuable main event references. We utilize the multilingual SBERT model (Reimers and Gurevych, 2020) that is pre-trained on numerous languages for vectorization, encoding the LM-generated event, the article title, and the lede sentences[1] to form a multi-chunk ME representation:

$$V_{\mathrm{ME}} = V_{\tilde{\mathrm{ME}}} \oplus V_{title} \oplus V_{lede} \tag{2}$$

where $\oplus$ is vector concatenation, ensuring chunk-wise representation comparison during clustering.

News clustering. Recent work by Zhang et al. (2024) has demonstrated the effectiveness of the dynamic clustering algorithm (DyClu) for large-scale social media text clustering, outperforming k-means, affinity propagation, and topic-based clustering approaches (MacQueen et al., 1967; Frey and Dueck, 2007; Llewellyn et al., 2014). Following Zhang et al. (2024), we optimize the following concave function to achieve dynamic thresholds during news clustering:

$$\gamma_t = \sqrt{\mathrm{K}_1 \times (\nu_t + \mathrm{K}_2)} \tag{3}$$

where K_1 and K_2 are hyper-parameters controlling the curve's steepness, and γ_t is the dynamic threshold calculated at iteration t given the number of articles ν_t. The iterative process terminates when increasing γ_t no longer increases cluster size. Further details regarding hyper-parameter tuning and the DyClu algorithm are shown in Appendix B.

3.2 MCMS

3.2.1 Data Sharpening Strategy

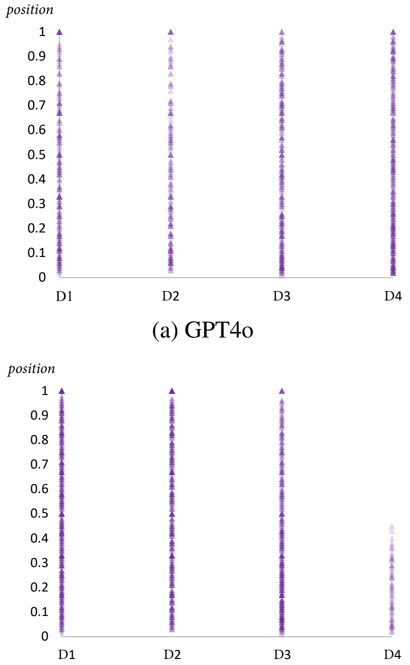
Figure 2 visualizes sentence referencing patterns of open- and closed-source LMs on the MCMS task. Only considering clusters of four articles, with sentence positions normalized to [0, 1] within each article, thus we can analyze both intra- and inter-article reference sentence distributions. We have two key observations: (1) GPT4o prioritizes sentences at the beginning of each article, with a particular emphasis on the latter two articles in each cluster. (2) Seallm-v2 (Nguyen et al., 2023) exhibits a heavy-tailed distribution, likely attributable to input sequence length limitations. The observed position biases in both LMs indicate a significant challenge for their practical application in MCMS scenarios.

This paper introduces a data-sharpening method to alleviate the above challenge. Given a multilingual article cluster, sentences can be grouped into various clusters based on their semantic similarity. Then we generate two sentence sequences, X and X', by sampling sentences from a single cluster and across different clusters, respectively. Assuming both sequences have the same length (m) and their word-level probability distributions are P and P', we argue that the words in X are likely more concentrated than those in X'. This implies that P is less uniform than P'. Compared to a uniform distribution U, we can therefore conclude:

$$\begin{aligned} H(P) &= \log m - D_{\mathrm{KL}}(P||U) \\ D_{\mathrm{KL}}(P||U) &> D_{\mathrm{KL}}(P'||U) \\ H(P) &< H(P') \end{aligned} \tag{4}$$

suggesting sampling sentences from various clusters encourages textual diversity with a high entropy value $H(P')$. With this inspiration, given a

[1] First 3 sentences of a news article serve as the lede.



(a) GPT4o

(b) SeaLLM-v2 (7B)

Figure 2: Sentence referencing patterns of LLMs.

news cluster containing N sentences, we apply the DyClu algorithm to group the sentences into distinct semantic clusters. The i-th cluster consists of c_i sentences. Given the selection ratio r, we select the $\max(1, r \times c_i)$ sentences closest to the centroid from each cluster. These selected sentences are then used to form a new document, denoted as X_r. Moreover, Figure 1 illustrates an interesting phenomenon: as the value of r shifts from 0.1 to 1, the information volume and entropy exhibit an inverse relationship. This suggests the importance of finding an optimal balance between the two. Inspired by this observation, we consider both information volume and entropy in our data-sharpening process, which is formulated as follows.

$$\begin{aligned} V(X_r) &= \xi_1 \frac{Z(X_r) - Z(X)_{\text{min}}}{Z(X)_{\text{max}} - Z(X)_{\text{min}}} + \xi_2 \\ \dot{r} &= \underset{r}{\text{argmax}} \left(\alpha H(X_r) + V(X_r)\right) \end{aligned} \tag{5}$$

where $Z(\cdot)$ denotes the application of *zlib* compression for data de-duplication, $V(X_r)$ represents the normalized information volume, ξ_1 is 0.9998, ξ_2 is 0.0001, α is a hyper-parameter scaling the entropy value for balancing purposes, and $\dot{r}$ signifies the resulting sampling ratio.

3.2.2 LM Fine-Tuning and Localization

Recent advancements in NLP have been significantly driven by the emergence of open-source LMs (Nguyen et al., 2023; Dubey et al., 2024; Yang et al., 2024; Team et al., 2024). In this work, we adopt SeaLLM-v2 (7B) (Nguyen et al., 2023) and Qwen2.5-Instruct (7B) (Yang et al., 2024) as foundational models for MCMS. For supervised fine-tuning, we use GPT4 to generate summaries for each article cluster, yielding a dataset D with 2K instances, 10% reserved for validation. Subsequently, we apply LoRA-based fine-tuning to the small LMs, guided by carefully crafted instructional prompts.

The above fine-tuned MCMS models can be inherently prone to knowledge bias stemming from its pre-training and fine-tuning data, which may become outdated when addressing open-domain knowledge in the ever-evolving news data of MCMS. Additionally, our analysis of system outputs reveals a lack of temporal sensitivity in the fine-tuned models, often leading to a mismatch between knowledge and timestamps. To mitigate this, we fine-tune a separate LoRA layer for temporal question answering generation (TQA-G) using 3K TQA instances by GPT4. Then we leverage this fine-tuned TQA-G model to build a larger dataset, D_{TQA}, with 400K TQA instances from local news. Finally, we jointly fine-tune the base model on MCMS and TQA to inject up-to-date knowledge extracted from recent local news articles.

$$\tau = \underset{\tau}{\text{argmin}} \Big(\mathbb{E}_{(x,y)\sim\{D;D_{\text{TQA}}\}}[L(\pi(x;\tau), y)] \Big) \tag{6}$$

Notably, we reuse the base LMs and fine-tune separate LoRA layers for multiple purposes (ME detection, TQA-G, TQA, and MCMS). This allows for dynamic knowledge updates, ensuring the final system remains sustainable in real-world MCMS applications. Further details on instruction and prompt designing can be found in Appendix C.

4 Data

4.1 SEASUMM-v1

This work presents the first dataset for evaluating the CLUST-MCMS pipeline task. It includes a manually created test set, focusing on news articles between 2022 and 2024 from Southeast Asia (SEA) in Malay (MS), Bahasa Indonesian (ID), Simplified Chinese (CN), and English (EN). To build gold-standard clusters, we first identify a sentence in each article that best encapsulates its main event. Using this approach, we select 30 hot topics, gather 473 related news articles, and manually extract the

Dataset	Sentence	Event	Article	Cluster
DEV	3,374	1,115	209	50
TEST	6,701	2,391	426	102

Table 1: Statistics for our built SEASUMM-v1 data.

main event from each article. The articles are then grouped manually based on the main events, resulting in 152 distinct news clusters. Furthermore, we introduce an additional 162 articles unrelated to the above events as noise to replicate real-world conditions. Prior work has demonstrated the effectiveness of protocol-guided prompting with GPT4 for generating silver-standard MCMS summaries, achieving high agreements of 0.96, 0.98, and 0.93 for *redundancy*, *omission*, and *conflict* (Ye et al., 2024). Inspired by this, we employ GPT4o to generate silver MCMS summaries for each cluster. These summaries undergo manual evaluation, achieving a high agreement of 91.67% regarding event coverage and faithfulness (0.5 points awarded per instance for meeting each defined goal). To support the coverage evaluation of MCMS in this work, we utilize GPT4o to identify event sentences related to manually identified ME for each article, obtaining 3,506 events in total. Manual validation of these events yields a high correctness agreement of 96.67%. More comprehensive statistics for our dataset are provided in Table 1.

4.2 GLOBESUMM

GLOBESUMM (Ye et al., 2024) introduces the first benchmark data[2] for the unified MCMS task. This dataset features high-quality summaries generated using protocol-guided prompting and spans 26 languages, including English, Italian, Arabic, etc., addressing diverse real-world needs. It is a good compensation for the MCMS datasets we constructed. In this paper, we adhere to their data-splitting setting, with 2,626 article clusters for training, 823 for validation, and 868 for testing.

5 Experiments

5.1 Metrics

For ENC, we use Normalized Mutual Information (NMI) to measure the agreement between the generated article cluster labels and the ground truth. Additionally, this work treats articles that do not belong to any cluster as a distinct category, referred to as isolated articles. To eliminate the impact of such on clustering performance, we also report a modified metric, denoted as NMIe, which excludes the 162 noise points (Subsection 4.1) in evaluation.

We evaluate MCMS using three metrics:

- **ROUGE:** We report standard F1-based ROUGE-1, ROUGE-2, and ROUGE-L scores, measuring the overlap co-occurrence of N-grams between the candidate and reference summaries.
- **Event Coverage (Eve-Cov):** This metric measures the extent to which the generated summary captures the manually extracted ME and related events. Specifically, a summary sentence $s_{i,k}$ is considered to cover an event $E_{i,j}$ in the i-th article cluster if their cosine similarity exceeds 0.75 (discussed in Subsection 5.4). On this basis, we calculate the micro-averaged P, R, and F1 scores across all summary sentences in the dataset as Eve-Cov performance.
- **Entity Faithfulness (Ent-Faith):** This metric evaluates the summary's coverage of source articles at the entity level. We define three entity types: **named** entities, **temporal** entities, and **quantitative** entities. Sentences lacking all three entity types are excluded from both summaries and source articles during evaluation. We calculate the accuracy value based on the exact match of all three entity types between the source and summary sentences.

For CLUST-MCMS, considering the mismatches between predicted and standard article clusters after ENC, we introduce a novel evaluation metric (**CLUST-ROUGE**) to align system outputs with ground truth as well as considering the N-gram correctness of generated summaries. First, we take a predicted summary as a "successful hit" if its cosine similarity with any standard summary exceeds 0.95. This allows us to compute the P, R, and F1 scores based on the number of successfully hit summaries. Furthermore, given the generated summaries and the corresponding hit standard summaries, we obtain their pair-wise ROUGE values, R-1, R-2, and R-L. On this basis, the ultimate CLUST-ROUGE scores are calculated as follows:

$$P = \frac{k}{M}; R = \frac{k}{N}; F1 = \frac{2 \times P \times R}{P + R} \qquad (7)$$
$$\text{CR-L} = F1 \times \text{R-L}$$

where k denotes the number of hit summaries, M and N denote the number of predicted and gold

[2] https://github.com/YYF-Tommy/GlobeSumm

Setting	NMI	Setting	NMIe
FaClu (2019)	86.15	FaClu (2019)	88.48
DyClu (2024)	87.31	DyClu (2024)	89.60
+ LEDE	87.53	+ LEDE	91.54
+ ME$_{llm}$	88.95	+ ME$_{llm}$	91.72
+ EWHO	90.75	+ EWHO	93.15
+ EWHERE	**90.83**	+ EWHERE	93.22
+ EWHEN ✗	88.92	+ EWHEN ✗	92.19
+ ETRIGGER ✗	89.86	+ ETRIGGER	**93.68**
+ EOUTCOME ✗	89.54	+ EOUTCOME ✗	92.89

Table 2: News clustering performance. "✗" denotes the setting is not applied in the following.

Method	R-1	R-2	R-L
–SEASUMM-v1–			
SeaLLM	48.35	26.48	31.11
SeaLLM*	55.09	30.63	35.84
SeaLLM* w/ DS	55.90	30.00	35.72
SeaLLM* w/ DS-enh	55.98	**30.88**	**36.42**
SeaLLM* w/ DS-enh†	54.68	29.58	34.85
GPT4	**56.45**	30.13	36.29
–GLOBESUMM–			
Qwen	37.90	9.11	16.36
Qwen*	37.34	7.80	15.08
Qwen* w/ DS	38.77	8.32	15.56
Qwen* w/ DS-enh	39.27	8.53	15.63
GPT3.5	35.74	11.35	18.40
GPT3.5 w/ DS	37.64	11.56	19.06
GPT3.5 w/ DS-enh	**39.95**	**12.73**	**19.74**
GPT4	35.61	11.03	18.06

Table 3: MCMS performance (ROUGE). "*" denotes instruct-tuning applied, "DS" and "DS-enh" denotes the basic and optimized data sharpening methods, respectively, and "†" denotes LM localization. Best results are bolded; second-best are underlined.

standard clusters, respectively. CR-1, CR-2, and CR-L refer to the resulting CLUST-ROUGE values. More details are provided in Appendix D.

5.2 Baseline Methods

For ENC, we employ the Fast Clustering (Reimers and Gurevych, 2019) (FaClu) and DyClu (Zhang et al., 2024) algorithms, along with the title and the top three sentences closest to the document vector (topic perspective) as ME references, to form our baseline approaches. For MCMS, we adopt the SeaLLM-v2 (Nguyen et al., 2023) as the baseline model, a compact language model pre-trained for SEA languages[3]. Additionally, we also report the performance of the large-scale state-of-the-art GPT4 model for reference. Since GLOBESUMM is specifically designed for the MCMS task, we follow (Ye et al., 2024) and evaluate our methods using GPT3.5 as well as the smaller Qwen2.5-Instruct model (Yang et al., 2024) for comparison.

5.3 Results

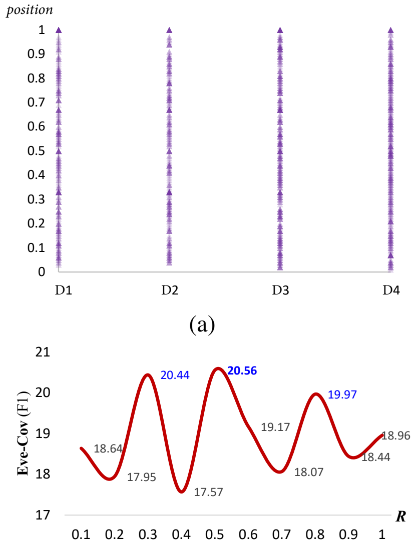
ENC performance. Table 2 reports the performance of our re-implemented FaClu and DyClu algorithms, along with an ablation study using various main-event reference settings. The table reports two groups of results, where NMI considers all articles for clustering evaluation, while NMIe excludes the effects of those noisy points and focuses more on the accuracy of the non-isolated articles. Comparing the FaClu and DyClu methods, we observe that employing the dynamic similarity thresholds (Zhang et al., 2024) enhances performance across both metrics. We also compare between eight different system modes. Among these, using the lede sentences (event perspective) as references outperforms the baseline method employing topic-level references, particularly for clustering the non-isolated articles. This indicates the significant difference between topic- and event-centric clustering. Furthermore, leveraging the LM-generated ME references can further boost the performance. As part of our methodology, we enrich ME by incorporating event attributes like WHO and WHERE as additional descriptive features. The results indicate that this approach leads to varied improvements across both metrics.

MCMS performance. Table 3 shows the MCMS performance of both small and large LMs when applied using our proposed approaches. Specifically, "DS" refers to using a fixed $\dot{r}$ in Eq. (5), selected based on the DEV set, while "DS-enh" denotes the optimized method, where Eq. (5) is fully applied. In our SEASUMM-v1 dataset, consistent with many recent studies, fine-tuning the smaller SeaLLM model using instruct-tuning data from GPT4 yields a significant performance improvement (+4.73 R-L). Under the DS setting, the performance improves for R-1 but decreases slightly for R-2 and R-L. However, when applying the enhanced data-sharpening (DS-enh) method, performance improves across all three metrics, highlighting the effectiveness of our strategies. Interestingly, fine-tuning the model with TQA data extracted from local news articles leads to a notable drop in ROUGE scores, underscoring the substantial impact of this action. We hypothesize that the fine-tuning of the 7B model on a large volume of TQA instances leads to mild forgetting in its text summarization capabilities. When comparing our fine-tuned small LMs with GPT4, the enhanced system achieves performance close to GPT4 on

[3] https://huggingface.co/SeaLLMs/SeaLLM-7B-v2

Method	P	R	F1	Faith$_{acc}$
–SEASUMM-v1–				
SeaLLM*	46.01	12.01	19.05	48.08
+ DS	46.69	12.01	19.11	51.49
+ DS-enh	46.94	12.50	19.75	53.68
+ DS-enh†	56.77	14.78	23.45	55.19
+ DS-enh†‡	**58.97**	**15.57**	**24.64**	**57.29**
GPT4	29.14	6.19	10.22	39.69
–GLOBESUMM–				
Qwen*	59.53	5.89	10.72	44.49
+ DS	64.53	6.27	11.43	47.41
+ DS-enh	64.13	**6.74**	**12.21**	45.45
GPT3.5	75.00	4.36	8.23	71.38
+ DS	74.83	4.44	8.38	**72.00**
+ DS-enh	**78.14**	4.90	9.22	70.97
GPT4	77.27	4.38	8.30	71.17

Table 4: MCMS results on Eve-Cov and Ent-Faith. "†" denotes LM localization and "‡" denotes the SFT data cleaning procedure. Due to the lack of localization data of GLOBESUMM, our localization experiments are only carried out on our contributed SEASUMM-v1 dataset annotated in this paper (see Section 7).



(a)

(b)

Figure 3: Event-level coverage analysis.

R-1 and surpasses GPT4 on R-2 and R-L. In the GLOBESUMM dataset, our proposed methods, when applied to Qwen and GPT3.5, yield varied yet consistent performance gains. Remarkably, the final performance significantly exceeds that of GPT4 across all three metrics.

We further evaluate the resulting systems using the proposed Eve-Cov and Ent-Faith metrics. The overall results in Table 4 demonstrate that applying the basic DS method to all three LMs consistently enhances their performance on the two metrics. This improvement underscores the clustering-filtering mode's effectiveness, enabling the models to focus on various sentence clusters with majority consensus and improving overall performance. However, the results exhibit significant variability when applying the enhanced DS method. For instance, DS-enh further boosts performance on both metrics for SEASUMM-v1, while for GLOBESUMM, it only improves Eve-Cov. Notably, fine-tuning the SeaLLM model with TQA for localization significantly improves performance on both metrics, contrasting sharply with the results in Table 3. This underscores the potential of our method to guide models in generating summaries with local styles. Furthermore, to further avoid the temporal mismatch between knowledge and timestamps, we clean the SFT data (source from GPT4) by requiring the model to strictly cite the facts in the original text to avoid knowledge abuse. The results (‡) show this simple action leads to further improvements on both metrics significantly. To our understanding, the inherent diversity of natural language allows for multiple valid silver-standard summaries, and those produced by GPT4o, while faithful and coherent, may reflect stylistic biases toward GPT4o itself. This poses challenges for traditional ROUGE-based evaluation, which prioritizes N-gram overlap and may yield results misaligned with Eve-Cov and Ent-Faith. Besides, the results reveal that GPT4 performs significantly worse in Southeast Asia scenarios, lagging behind our final system by over 10 points on both metrics. In light of this, we conduct a case study in Section 5.4 to further discuss it.

CLUST-MCMS performance. The pipeline system takes articles containing noisy points as inputs to the clustering module, with the predicted clusters further input into the MCMS module. The results in Table 5 demonstrate that the utilized clustering and MCMS methods consistently enhance the overall performance on our CLUST-ROUGE metrics. Besides, applying the localization method leads to a performance drop, aligning with the previous observations in Table 3. As we can see, the pipeline performance is significantly influenced by both the F1 score of correctly hit summaries, as well as the pair-wise ROUGE values. For example, applying DyClu (Line 2) yields a substantial F1 increase, yet the R-2 score drops markedly. This trade-off ultimately leads to notable improvements on CR-1 and CR-L despite the decrease on CR-2. These findings demonstrate the robustness of the proposed CLUST-ROUGE metric to some extent.

Method	P	R	F1	R-1	R-2	R-L	CR-1	CR-2	CR-L
FaClu w/ McMs-basic	85.54	69.61	76.76	57.55	32.44	36.22	44.18	24.90	27.80
DyClu w/ McMs-basic	83.70	75.49	79.38	57.20	31.31	36.42	45.41	24.86	28.91
DyClu w/ McMs-enh	84.78	76.47	**80.41**	57.66	31.29	**36.64**	**46.37**	**25.16**	**29.47**
DyClu w/ McMs-enh†	82.61	74.51	78.35	56.87	30.93	36.60	44.56	24.23	28.68

Table 5: Event-centric new clustering and McMs pipeline performance. "McMs-basic" denotes taking an entire news cluster as input to the fine-tuned SeaLLM model. "McMs-enh" denotes the system with DS-enh applied.

Example news cluster
Aspects: China's peaceful reunification efforts; Military activities and legal basis for force; "One country, two systems" proposal, and its rejection; Taiwan's stance and sovereignty; skepticism and public opinion in Taiwan; Historic meetings and cross-strait diplomacy.
Summary by GPT4
On September 21, 2022, tensions between China and Taiwan were highlighted as Chinese officials reiterated their commitment to the "reunification" of Taiwan with the mainland. China's Taiwan Affairs Office, represented by spokespeople such as Mr. Ma Xiaoguang and Mr. Qiu Kaiming, emphasized China's strategic patience and confidence in achieving unification, despite skepticism from Taiwan. They portrayed a belief that more Taiwanese are realizing that their future lies in unification, a claim contradicted by recent polls indicating strong opposition to unification and perceptions of Beijing as unfriendly. Following Nancy Pelosi's visit to Taiwan, which marked the first such visit by a sitting US House Speaker in 25 years, China conducted significant military drills, including the firing of ballistic missiles over Taiwan.
Summary by our system
On Wednesday, September 21, 2022, China's Taiwan Affairs Office spokesperson, Ma Xiaoguang, stated during a news conference in Beijing that China is willing to make the utmost effort to strive for a peaceful "reunification" with Taiwan. This statement was made following weeks of military maneuvers and war games by China near Taiwan, which began after a visit by United States House Speaker Nancy Pelosi to Taipei in August. Ma emphasized China's determination to safeguard its territory and asserted that the reunification of Taiwan with the mainland is an inevitable historical trend. He also mentioned that Taiwan could have a "social system different from the mainland" under the condition of respecting their way of life, including religious freedoms. However, China's proposal of a "one country, two systems" model, similar to the formula used for Hong Kong's return to Chinese rule in 1997, has been rejected by all mainstream Taiwanese political parties and has almost no public support, particularly after the imposition of a national security law on Hong Kong in 2020. China has not renounced the use of force to bring Taiwan under its control and has a legal basis for military action against Taiwan if it secedes or appears to be doing so.

Table 6: Case study. We represent the news cluster with abstracted aspects for space limitation (see Appendix E).

5.4 Analysis and Discussion

Correlation between input entropy/volume and output coverage. Our experiments demonstrate that filtering input sentences using the DS and DS-enh frameworks improves event-level coverage. To better understand the results, Figure 3 (a) illustrates the distribution of sentence selection in DS, which appears more evenly spread across all article cluster positions when compared with the LMs' distributions shown in Figure 2. Additionally, analysis of sentence selection ratios on the DEV set reveals three coverage score peaks in Figure 3 (b), with the highest score achieved at a 0.5 ratio, aligning closely with the central positions. When cross-referenced with entropy and volume data in Figure 1, the results suggest that optimizing the trade-off between entropy and volume makes sense for McMs. We argue this approach can yield better results for large-scale summarization tasks.

Case study. The previous results demonstrate a significant performance gap between GPT4 and our methods with smaller LMs. To further investigate this, we conduct a case study. Table 6 presents an article cluster with six manually abstracted aspects, along with the summaries generated by GPT4 and ours. We manually evaluate how well each summary covers the aspects, assigning a score from 0 to 5 for each aspect, obtaining [3, 3, 0, 3, 4, 3] for GPT4 and [5, 5, 5, 4, 4, 3] for ours, ordered according to the aspects. The results indicate that our method generates a more comprehensive and faithful summary, covering most aspects in greater detail, whereas GPT4 performs worse on the first four aspects. As previously discussed in Figure 2, one potential reason for this discrepancy is the positional bias hidden within LLMs, which may cause it to allocate disproportionate attention to the aspect it deems most important, resulting in less focus on the other aspects.

Human agreement. Unlike Ent-Faith, which is calculated based on exact word-level matches with minimal subjectivity, the Eve-Cov metric may be influenced by imperfect sentence representations and the choice of similarity threshold, introducing some subjectivity. To assess its reliability, we conduct a manual evaluation of Eve-Cov, comparing the agreements among human annotators, GPT4o evaluation, and our proposed Eve-Cov metric. The overall results are presented in Table 7. It show that our metric demonstrates a +2.46 points improvement in agreement compared to GPT4o, highlighting the reliability of our introduced evaluation metric.

Metric	Agreement
GPT4o	82.27
Eve-Cov	**84.73**

Table 7: Human agreement on Eve-Cov.

6 Conclusion

This work introduced a novel task of event-centric news cluster summarization. Our investigation into the impact of varied ME references on clustering, coupled with our proposed summarization framework balancing information volume and entropy, demonstrated significant performance gains. Furthermore, the joint fine-tuning strategy for LM localization yielded substantial improvements in capturing localized context. These findings contribute valuable insights into leveraging LM capabilities for enhanced cross-lingual and localized news analysis, opening promising avenues for future research.

7 Limitations

We utilize the open-source GLOBESUMM dataset and conduct supplementary experiments to demonstrate the effectiveness of our proposed methods. However, we encountered challenges in the LM localization experiments due to the GLOBESUMM dataset's extensive coverage of up to 26 languages, which is primarily designed to test the global capabilities of models. The localization task for such a large, diverse set of languages proved to be quite complex. Furthermore, obtaining sufficient, up-to-date localization data for all 26 languages simultaneously is challenging. In this situation, using small-scale data for knowledge editing does not yield meaningful results, which leads to a lack of localization practice for GLOBESUMM in this paper. In addition, our LM localization strategy in this paper focuses more on language and news areas with fewer resources, where language models are still underdeveloped and could benefit most from such localization efforts.

8 Acknowledgments

We thank Nattadaporn Lertcheva, Nabilah Binte Md Johan, Wang Xi, and other experienced researchers in our team for their insightful discussions and contributions to the data construction.

References

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.

James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. OpenAsp: A benchmark for multi-document open aspect-based summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1991, Singapore. Association for Computational Linguistics.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Salvatore Carta, Sergio Consoli, Luca Piras, Alessandro Sebastian Podda, and Diego Reforgiato Recupero. 2021. Event detection in finance using hierarchical clustering algorithms on news and tweets. *PeerJ Computer Science*, 7:e438.

Wang Chen, Piji Li, and Irwin King. 2021. A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy. 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 404–414, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

DeepSeek, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ahmed Elhady, Khaled Elsayed, Eneko Agirre, and Mikel Artetxe. 2024. Improving factuality in clinical abstractive multi-document summarization by guided continued pre-training. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 755–761, Mexico City, Mexico. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. MSAMSum: Towards benchmarking multi-lingual dialogue summarization. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 1–12, Dublin, Ireland. Association for Computational Linguistics.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

Poonam Goyal, Prerna Kaushik, Pranjal Gupta, Dev Vashisth, Shavak Agarwal, and Navneet Goyal. 2019. Multilevel event detection, storyline generation, and summarization for tweet streams. *IEEE Transactions on Computational Social Systems*, 7(1):8–23.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593, Mexico City, Mexico. Association for Computational Linguistics.

David A Huffman. 1952. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024. Polarity calibration for opinion summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5211–5224, Mexico City, Mexico. Association for Computational Linguistics.

Quanzhi Li, Yang Chao, Dong Li, Yao Lu, and Chi Zhang. 2022. Event detection from social media stream: methods, datasets and opportunities. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3509–3516. IEEE.

Zehang Lin, Jiayuan Xie, and Qing Li. 2024. Multimodal news event detection with external knowledge. *Information Processing & Management*, 61(3):103697.

Mathis Linger and Mhamed Hajaiej. 2020. Batch clustering for multilingual news streaming. *arXiv preprint arXiv:2004.08123*.

Zheng Liu, Yu Zhang, Yimeng Li, and Chaomurilige. 2023. Key news event detection and event context using graphic convolution, clustering, and summarizing methods. *Applied Sciences*, 13(9):5510.

Clare Llewellyn, Claire Grover, and Jon Oberlander. 2014. Summarizing newspaper comments. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 599–602.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020. Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1737–1751, Online. Association for Computational Linguistics.

Sebastião Miranda, Artūrs Znotiņš, Shay B. Cohen, and Guntis Barzdins. 2018. Multilingual clustering of streaming news. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4535–4544, Brussels, Belgium. Association for Computational Linguistics.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Olubusayo Olabisi and Ameeta Agrawal. 2024a. Understanding position bias effects on fairness in social multi-document summarization. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 117–129, Mexico City, Mexico. Association for Computational Linguistics.

Olubusayo Olabisi and Ameeta Agrawal. 2024b. Understanding position bias effects on fairness in social multi-document summarization. *arXiv preprint arXiv:2405.01790*.

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2025–2031, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph Peper, Wenzhao Qiu, and Lu Wang. 2024. PELMS: Pre-training for effective low-shot multi-document summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7652–7674, Mexico City, Mexico. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

João Santos, Afonso Mendes, and Sebastião Miranda. 2022. Simplifying multilingual news clustering through projection from a shared space. *arXiv preprint arXiv:2204.13418*.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Adane Nega Tarekegn, Fazle Rabbi, and Bjørnar Tessem. 2024. Large language model enhanced clustering for news event detection. *arXiv preprint arXiv:2406.10552*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. ClidSum: A benchmark dataset for cross-lingual dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Towards unifying multi-lingual and cross-lingual summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15127–15143, Toronto, Canada. Association for Computational Linguistics.

Lin Wu, Rui Li, and Wong-Hing Lam. 2023. Research on multilingual news clustering based on cross-language word embeddings. *arXiv preprint arXiv:2305.18880*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 118–127, Lisbon, Portugal. Association for Computational Linguistics.

Yangfan Ye, Xiachong Feng, Xiaocheng Feng, Weitao Ma, Libo Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024. GlobeSumm: A challenging benchmark towards unifying multi-lingual, cross-lingual and multi-document news summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10803–10821, Miami, Florida, USA. Association for Computational Linguistics.

Jianmin Zhang, Jiwei Tan, and Xiaojun Wan. 2018. Adapting neural single-document summarization model for abstractive multi-document summarization: A pilot study. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 381–390, Tilburg University, The Netherlands. Association for Computational Linguistics.

Longyin Zhang, Bowei Zou, Jacintha Yi, and AiTi Aw. 2024. Comprehensive abstractive comment summarization with dynamic clustering and chain of thought. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2884–2896, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xin Zheng, Aixin Sun, Jing Li, and Karthik Muthuswamy. 2019. Subtopic-driven multi-document summarization. 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)*, pages 3153–3162, Hong Kong, China. Association for Computational Linguistics.

A Information Entropy and Volume Visualization

We visualize the information entropy and volume values in Figure 1 to analyze and motivate our research. For each article cluster, we apply the clustering-based sampling method ten times, varying the sampling ratio from 0.1 to 1. For a given ratio r, we obtain a set of sampled sentences and calculate their information entropy and volume, e_r and v_r. To effectively visualize the distribution of e_r and v_r within each independent article cluster, we normalize them to the range [0, 1].

$$\begin{aligned}\text{NormEnt} &= \lambda_1 \frac{(e_r - e_{min})}{e_{max} - e_{min}} + \lambda_2 \\ \text{NormVol} &= \lambda_1 \frac{(v_r - v_{min})}{v_{max} - v_{min}} + \lambda_2\end{aligned} \tag{8}$$

where λ_1 = 0.9998 and λ_2 = 0.0001.

B Experimental Settings

The FaClu algorithm maintains a similarity threshold to determine if an article belongs to a cluster or not. We select the threshold based on our DEV data, the parameter selection process is detailed in Table 8.

Algorithm 1 - DyClu

```
Input: $N$ vectorized news articles $C$
Output: a list of article clusters $G$
Initialization: threshold: $\gamma$, cluster size: $\nu$, maximum threshold: $thr_{max} = 0.9$
Begin
for $t$-th article $c_t$ in $C$ do
    scores ← pairwise_cos_sim($c_t$, $C$)
    scores-k← scores.top-k($\nu$)
    $\gamma' \leftarrow \gamma$
    while scores-k[-1] > $\gamma'$ and $\nu$ < $N$ do
        $\nu \leftarrow \text{Min}(N, \nu + \Delta)$
        $\gamma' \leftarrow \text{Min}(\sqrt{\text{K}_1 \times (\nu + \text{K}_2)}, thr_{max})$
        scores-k← scores.top-k($\nu$)
    end while
    $G_t \leftarrow []$
    for $s_i$ in scores-k do
        $G_t \leftarrow G_t \cup \{c_i\}$ when $s_i \geq \gamma'$
    end for
    $G \leftarrow G \cup \{G_t\}$
    $R_t = \Sigma_j^n \text{cosine}(c_j, c_t)$
end for
Rank the $N$ clusters in $G$ based on $R_{1...N}$
End
```

θ	**NMI**
0.5	78.70
0.52	**79.30**
0.54	78.93
0.56	76.64
0.58	74.62
0.60	70.15

Table 8: FaClu similarity threshold selection.

The DyClu method introduces two parameters, K_1 and K_2, as shown in Eq. (3), to shape the curve and dynamically adjust the similarity thresholds. To define this curve, we anchor it at two specific points in the 2D space based on performance on our DEV set. Specifically, we select two cluster numbers (2 and 12 in this paper) on the X-axis and determine the corresponding similarity thresholds, γ_1 and γ_2, that yield the best DEV performance. The parameter selection process is detailed in Table 9. More details about our used DyClu clustering can be found in Algorithm 1.

We use a fixed sentence sampling ratio for the basic data-sharpening method (i.e., DS in Table 3). The parameter selection results for SEASUMM-v1

γ_1	γ_2	**NMI**
0.5	$\gamma_1 + [0.02, \dots, 0.2]$	**78.98**, γ_2=**0.54**
0.52	$\gamma_1 + [0.02, \dots, 0.2]$	78.60, γ_2=0.54
0.54	$\gamma_1 + [0.02, \dots, 0.2]$	78.84, γ_2=0.56
0.56	$\gamma_1 + [0.02, \dots, 0.2]$	75.28, γ_2=0.60
0.58	$\gamma_1 + [0.02, \dots, 0.2]$	74.52, γ_2=0.60
0.60	$\gamma_1 + [0.02, \dots, 0.2]$	69.74, γ_2=0.66

Table 9: DyClu curve selection.

r-A	**F1**	**r-B**	**F1**
0.1	18.64	0.1	10.53
0.2	17.95	0.2	9.83
0.3	20.44	0.3	10.35
0.4	17.57	**0.4**	**11.83**
0.5	**20.56**	0.5	10.92
0.6	19.17	0.6	10.93
0.7	18.07	0.7	11.45
0.8	19.97	0.8	11.74
0.9	18.44	0.9	11.72
1.0	18.96	1.0	11.64

Table 10: Ratio selection for basic DS.

(r-A) and GLOBESUMM (r-B) are presented in Table 10. For the enhanced data-sharpening method, the hyper-parameter α in Eq. (5) is introduced to balance the two types of information. The parameter selection process is detailed in Table 11, where α-A and α-B correspond to the SEASUMM-v1 and GLOBESUMM datasets, respectively. Parameters are selected based on the Eve-Cov scores to minimize uncertainty in the silver-standard summaries, ensuring alignment with events in source articles.

All manual annotations were conducted in-house by the authors and five additional experienced researchers from our team, and the data is distributed according to the native language habits of the annotators. All systems were implemented using the PyTorch framework and trained on two A40 GPU cards. Model selection was based on the loss values on the validation set. The reported results were obtained through multiple experiments, with minor fluctuations that do not affect the overall conclusions. All our used pre-trained models are available for research purposes. *zlib* can be found in `https://docs.python.org/3/library/zlib.html`.

C Prompt/Instruction Designing

For ME generation:

> “**Article:**\n**[news content]**\n\n **Instructions:**\nPlease review the article, determine the main event of the article, and extract the event trigger, arguments (Who, Where, When), and outcome in English. Return the main event information in the format of ‘Event: ... | Trigger: ... | Who: ... | Where: ... | When: ... | Outcome: ...’, making the summary more accurately reflects the content and details of the article.\n\n**Main event information:**”

α**-A**	**F1**	α**-B**	**F1**
0.1	20.80	0.1	12.00
0.2	19.99	0.2	11.43
0.3	20.80	**0.3**	**12.05**
0.4	20.37	0.4	11.95
0.5	20.01	0.5	11.79
0.6	20.74	0.6	11.69
0.7	20.03	0.7	11.51
0.8	**21.65**	0.8	11.41
0.9	20.13	0.9	11.84
1.0	20.82	1.0	10.95

Table 11: Parameter tuning for enhanced DS.

For MCMS:

> “The following are sentences from articles in languages like English, Chinese, or Malay, each sentence is attached with a unique sentence ID. Please identify events among these sentences and summarize the events into a fluent paragraph. The articles are:\n**[news content]**\nSummarize the articles in English and ensure the summary is faithful, concise with a moderate length, and covers the main points. Ensure all proper nouns are fully presented, including person names, organization names, locations, etc. If the event has a clear occurrence time in the original text, please formally reflect the timestamp in the summary. Return the summary as a piece of text.”

For TQA-generation:

> “Below is a news published on **[news publication date]**, generate a temporal question-answer pair based on the context of the following article and the answer should sourced from the article.\n\n**[news content]**\n\nPlease return the question-answer pair in the format of ‘Question: ... | Answer: ...’ ”

For TQA:

> “Below is a news published on **[news publication date]**, generate a temporal question-answer pair based on the

context of the following article and the answer should sourced from the article.\n\n**[news content]**\n\nQuestion: **[question content]** | Answer:”

D Evaluation Details

As outlined in the main body, we assess the generated summaries based on entity-level faithfulness. The evaluation focuses on three categories of entities: named entities (PERSON, ORG, GPE, and LOC), temporal entities (DATE), and quantitative entities (CARDINAL, QUANTITY, and MONEY). We obtain the ground-truth entities in two stages, first using the google translation API to convert source articles into English, then extracting entities from source and summary sentences using spacy[4]. We employ INSTRUCTOR (Su et al., 2023)[5] to represent summaries for pipeline evaluation. It is a task-aware embedding model fine-tuned on 330 tasks using instruction-based methods. We adopt the instruction “Represent the statement:” as described in the original paper to represent each summary for similarity calculations.

E Supplement for Case Study

Table 12 serves as a supplement to Table 6.

[4] https://pypi.org/project/spacy/

[5] https://instructor-embedding.github.io/

Example news cluster
Events: [‘BEIJING – China is willing to make the utmost effort to strive for a peaceful “reunification” with Taiwan, a government spokesman said on Wednesday, following weeks of Chinese military manoeuvres and war games near the self-ruled island.’, ‘Mr Ma Xiaoguang, a spokesman for China’s Taiwan Affairs Office, issued the statement at a news conference in Beijing.’, ‘"The motherland must be reunified and will inevitably be reunified," he said.’, ‘"Chinaś determination to safeguard its territory is unwavering."’, "Taiwan’s government said on Wednesday that Taiwan will never allow China to “meddle” in its future.", ‘Taiwan’s Mainland Affairs Council said the island’s future was up to its 23 million people to decide.’, ‘Mr Ma’s comments came days after President Joe Biden said US military forces will defend Taiwan if there is “an unprecedented attack”.’, ‘China lodged a formal complaint with the US in response, pointing out that Mr Biden’s comments send a “seriously wrong signal” to separatist forces in Taiwan.’, ‘China has proposed a “one country, two systems” model for Taiwan, similar to the formula under which the former British colony of Hong Kong returned to Chinese rule in 1997.’, ‘“The Taiwanese people have already clearly rejected it,” the Mainland Affairs Council said.’, "China has refused to talk to Taiwan’s President Tsai Ing-wen since she first took office in 2016, viewing her as a separatist.", ‘She has repeatedly offered to talk on the basis of equality and mutual respect.’, ‘But Ms Tsai’s predecessor, Mr Ma Ying-jeou, held a landmark meeting with Chinese President Xi Jinping in Singapore in 2015.’, "Speaking at the same news conference, Mr Qiu Kaiming, head of the research department at the Communist Party of China’s Taiwan Work Office, said the Xi-Ma meeting showed their “strategic flexibility” towards Taiwan.", ‘That “showed the world that Chinese people on both sides of the (Taiwan) Strait are absolutely wise and capable enough of solving our own problems”, he said.’, ‘Taiwan’s government says that as the island has never been ruled by the People’s Republic of China, Beijing’s sovereignty claims are void.’, ‘BEIJING - China said it has the patience to someday bring Taiwan under its control, partly because "compatriots" there want it to happen - a view that contrasts with polling showing sceptical views of Beijing.’, ‘"With regard to resolving the Taiwan question and realising the complete unification of China, we have strategic composure and historic patience, and we are also full of confidence," Mr Qiu Kaiming, an official in a Chinese government department that handles ties with the island, said at a Wednesday news briefing.’, ‘"More and more Taiwan compatriots realise the future of Taiwan lies in the national unification," said Mr Qiu, who was speaking at a briefing in Beijing held by the Taiwan Affairs Office (TAO) to recap ties over the past decade.’, ‘He added that "the vast majority of Taiwan people oppose independence".’, ‘China regularly touts the measures it takes to try to win over the 23 million people of Taiwan, such as introducing policies to attract businesses and students.’, ‘TAO spokesman Ma Xiaoguang sidestepped a question at the briefing on Wednesday about Chinaś timetable for taking control of Taiwan, saying only that "itś a historic trend that no one will be able to stop".’, ‘BEIJING: China is willing to make the utmost effort to strive for a peaceful "reunification" with Taiwan, a Chinese government spokesperson said on Wednesday (Sep 21), following weeks of military manoeuvres and war games by Beijing near the island.’, ‘China claims democratically-governed Taiwan as its own territory.’, "Taiwan’s government rejects China’s sovereignty claims and says only the island’s people can decide their future.", ‘China has been carrying out drills near Taiwan since early last month after United States House Speaker Nancy Pelosi visited Taipei, including firing missiles into waters near the island.’, ‘Ma Xiaoguang, a spokesperson for Chinaś Taiwan Affairs Office, told a news conference in Beijing that China was willing to make the greatest efforts to achieve peaceful "reunification".’, ‘"The motherland must be reunified and will inevitably be reunified," Ma said.’, "China’s determination to safeguard its territory is unwavering, he added.", ‘China has proposed a "one country, two systems" model for Taiwan, similar to the formula under which the former British colony of Hong Kong returned to Chinese rule in 1997.’, ‘Ma said Taiwan could have a "social system different from the mainland" that ensured their way of life was respected, including religious freedoms, but that was "under the precondition of ensuring national sovereignty, security, and development interests".’, ‘All mainstream Taiwanese political parties have rejected that proposal and it has almost no public support, according to opinion polls, especially after Beijing imposed a national security law on Hong Kong in 2020 after the city was rocked by sometimes violent anti-government and anti-China protests.’, ‘China has also never renounced the use of force to bring Taiwan under its control, and in 2005 passed a law giving the country the legal basis for military action against Taiwan if it secedes or seems about to.’, ‘China has refused to talk to Taiwan President Tsai Ing-wen since she first took office in 2016, believing she is a separatist.’, ‘She has repeatedly offered to talk on the basis of equality and mutual respect.’, "But Tsai’s predecessor Ma Ying-jeou held a landmark meeting with Chinese President Xi Jinping in Singapore in 2015.", ‘Speaking at the same news conference, Qiu Kaiming, head of the research department at the partyś Taiwan Work Office, said the Xi-Ma meeting showed their "strategic flexibility" towards Taiwan.’, ‘That "showed the world that Chinese people on both sides of the Strait are absolutely wise and capable enough of solving our own problems", he added.’]

Table 12: The complete content of events manually detected from the news clusters in the case study section.