MRGSEM-Sum: An Unsupervised Multi-document Summarization Framework based on Multi-Relational Graphs and Structural Entropy Minimization

Anonymous ACL submission

Abstract

Multi-document summarization faces challenges such as complex inter-document relationships and information redundancy. Graph clustering provides an effective paradigm for addressing these challenges: it models the complex relationships among documents through graph structures and reduces information redundancy via clustering, achieving substantial progress in prior research. However, existing methods often consider only single-relational graphs and require the number of clusters to be predefined, which limits their ability to fully capture rich relational information and weakens their effectiveness in adaptively partitioning sentence groups to reduce redundancy. To overcome these limitations, we propose MRGSEM-Sum, an unsupervised multi-document summarization framework based on multi-relational graphs and structural entropy minimization. Specifically, we construct a multi-relational graph that integrates semantic and discourse relations between sentences, comprehensively modeling the intricate and dynamic connections among sentences across documents. We then apply a two-dimensional structural entropy minimization algorithm for clustering, automatically determining the optimal number of clusters and effectively organizing sentences into coherent groups. Finally, we introduce a position-aware compression mechanism to distill each cluster, generating concise and informative summaries. Extensive experiments on four benchmark datasets (Multi-News, DUC-2004, PubMed, and WikiSum) demonstrate that our approach consistently outperforms previous unsupervised methods and, in several cases, achieves performance comparable to supervised models and large language models.

1 Introduction

Multi-Document Summarization (MDS) is designed to create compact and informative summaries from collections of documents that revolve around specific subjects, facilitating users in efficiently gaining essential information. (Ma et al., 2022; Roy and Kundu, 2023; Li et al., 2023). However, due to the large number of documents, the relationships between documents are highly complex, and there is also the issue of information redundancy among documents, leading to suboptimal results with existing methods.

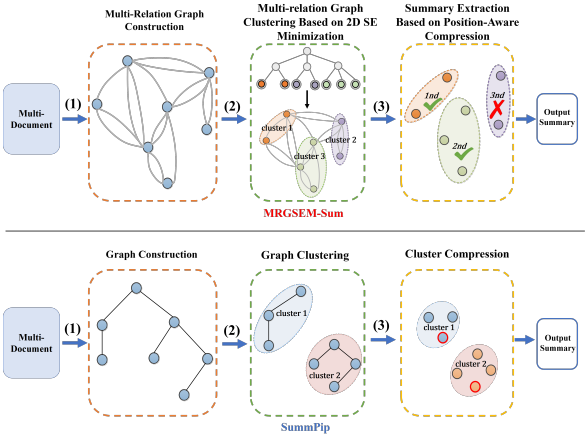

Figure 1: Model architecture comparison between MRGSEM-Sum and the existing graph-based multi-document summarization method (i.e., Sumpip)

In response, researchers have conducted in-depth studies. Among them, unsupervised methods eliminate the need for annotated data and instead leverage the intrinsic structure and semantic relationships within document clusters to automatically generate summaries. This not only reduces data acquisition costs but also provides stronger generalization capabilities, enabling these methods to flexibly adapt to various topics and domains (Vogler et al., 2022). Therefore, unsupervised MDS methods have become a focal point in recent MDS research. Existing unsupervised methods can be categorized into autoencoders (Vogler et al., 2022), ranking-based approaches (Lamsiyah et al., 2021; Alambo et al., 2020), and graph clustering methods Zhao et al. (2020); Alami et al. (2021). While

autoencoders can effectively retain detailed document information, they often struggle to filter out the core content required for summarization and face challenges with long input sequences (Zhao et al., 2020). Ranking-based methods, while able to identify salient sentences, are usually grounded in rigid prior assumptions, such as semantic relevance and diversity, which may not always hold in practice. In contrast, graph clustering methods, such as Sumpip (Zhao et al., 2020), model the complex relationships between documents by constructing a graph of relationships between sentences and then reducing redundancy among documents through clustering. This method can straightforwardly address the complexities of relationships and information redundancy in MDS and has become a current research hotspot. However, existing methods face two limitations: **First, existing graph structures are typically limited to modeling single-relation connections, constraining their ability to capture the complex and heterogeneous relationships that naturally arise in multi-document scenarios, such as semantic, discourse, and pragmatic links. Second, traditional clustering frameworks often require the number of clusters to be predefined, increasing model complexity and inflexibility.**

To address these limitations, we propose an unsupervised multi-document summarization framework based on multi-relational graphs and structural entropy minimization, named MRGSEM-Sum. Specifically, we first construct a multi-relational graph that captures semantic and discourse relationships among sentences. We then utilize a two-dimensional (2D) structural entropy (SE) minimization algorithm to cluster the sentence graph in an unsupervised manner, automatically determining the optimal number of clusters and grouping highly relevant sentences together. Finally, we introduce a position-aware compression strategy to further refine and condense each cluster, resulting in coherent and comprehensive summaries. A visual comparison of the model architectures of MRGSEM-Sum and the existing graph-based multi-document summarization method (i.e., Sumpip) is illustrated in Figure 1. In summary, our key contributions are as follows:

- We propose a novel unsupervised multi-document summarization framework that addresses the complexities of relationships and information redundancy in MDS.
- We design a multi-relational graph construction method, combine with the 2D SE minimization clustering approach and position-aware compression mechanism, to capture the intrinsic complex relationships between documents, eliminate information redundancy, and enhance the quality of summaries.
- We conduct extensive experiments on four benchmark MDS datasets, including Multi-News, DUC-2004, PubMed, and WikiSum. The results demonstrate that MRGSEM-Sum outperforms state-of-the-art unsupervised baselines significantly in both automatic metrics and human evaluation, showing competitiveness with supervised methods and large language models.

2 Methodology

As shown in Figure 2, MRGSEM-Sum consists of three modules.

2.1 A. Multi-relational Graph Construction

To capture the complex and diverse relationships between documents in the input document cluster $D = \{d_1, d_2, ..., d_n\}$, we first construct a multi-relational graph $G = \{V, E_{r_1}, E_{r_2}\}$ utilizing both semantic relationship and discourse relationship, as illustrated in Figure 2A. The two types of relationships, approached from different perspectives, complement each other, thereby modeling the complex relationships between documents.

Specifically, semantic relationship based edges E_{r1} are determined by calculating the cosine similarity between frequency-inverse document frequency (TF-IDF) representations of sentences. The weight of these edges reflects the similarity at the content level, providing a quantitative measure of the potential associations between sentences. Discourse relationship based edges E_{r2} are established based on the method proposed by Cristensen et al.(Christensen et al., 2013). We consider features such as discourse markers, coreference, and entity linking to establish discourse relationships between sentences. These edges capture not only the structural connections of sentences within the documents but also reveal their logical and rhetorical relationships. The nodes V are sentences from the input document cluster, we select a pre-trained sentence embedding model SBERT to initialize the representation of nodes, thereby obtaining high-dimensional node representations X.

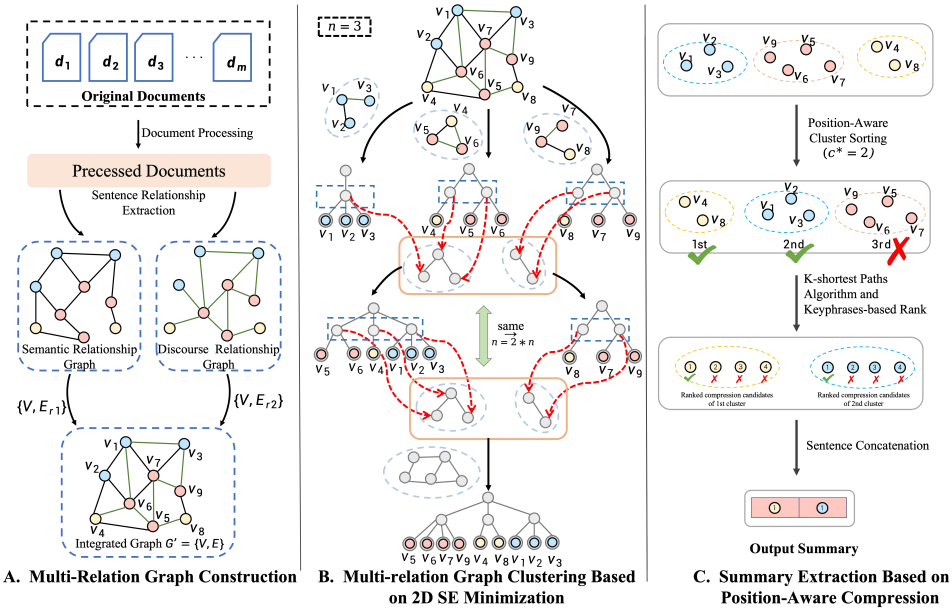

Figure 2: The framework of MRGSEM-Sum. A. Multi-relational Graph Construction. First, a multi-relational graph is constructed with sentences from multiple documents as nodes, based on sentence semantic similarity and discourse relations across the documents. B. Multi-relation Graph Clustering Based on 2D SE Minimization. Based on the constructed multi-relational graph and following the principle of SE minimization, an optimal 2D encoding tree is established. Sentences are assigned to different subtrees on the encoding tree, thereby partitioning the original multi-relational graph into subgraphs. C. Summary Extraction Based on Position-Aware Compression. The subgraphs of sentences are then compressed using a position-aware compression method to generate the final summary text.

2.2 Multi-relation Graph Clustering Based on 2D SE Minimization

Existing unsupervised summarization methods often utilize clustering models on document relation graphs to aggregate similar sentences and reduce redundancy (Alami et al., 2021; Wang et al., 2021). However, these approaches typically require predefined cluster numbers, which is impractical for real-world applications. To address this limitation, we propose an adaptive clustering method based on 2D SE minimization that automatically determines the optimal cluster structure, as illustrated in Figure 2B.

Since the 2D SE cannot be directly applied to multi-relational graphs, we need to first merge the relationships within the multi-relational graphs to form an integrated graph. The merging process is as follows:

$$A'[i][j] = \begin{cases} \max(A_{E_{r_1}}[i][j], A_{E_{r_2}}[i][j]) & \text{if } A_{E_{r_m}}[i][j] \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $A^{'}$ represents the adjacency matrix of the integrated graph $G^{'} = \{V, E\}$, and $A^{'}[i][j]$ indicates the weight of the edge from node i to node j. The elements $A_{E_{r1}}[i][j]$ and $A_{E_{r2}}[i][j]$ correspond to the weights of the edges between node i and j under the relationship types E_{r1} and E_{r2}, respectively.

Then, we apply a clustering algorithm based on 2D SE minimization to the integrated graph $G' = \{V, E\}$, as shown in **Algorithm 1**. Specifically, each sentence initially resides in its own cluster (line 1). Subsequently, these clusters are combined into subsets of size n (line 3), and each subset is transformed into a subgraph $G'_s = \{V', E'\}$ (lines 5-7). Then, each subgraph is converted into an encoding tree T' (lines 9-14). The constructed encoding tree adheres to the following definition (Li and Pan, 2016):

1. Let T' be a encoding tree structure over the set V', where each node α in T' is linked to a subset $T'_\alpha \subseteq V'$. The root node is denoted by λ, contains the entire set, i.e., $T'_\lambda = V'$. In contrast, each leaf node γ corresponds uniquely to a single node from V', such that $T'_\gamma = \{v\}$ for some $v \in V'$.

2. For any node α in T' with children $\beta_1, \ldots, \beta_k$, the sets $T'_{\beta_1}, \ldots, T'_{\beta_k}$ form a partition of T'_α.

3. For every node α in T', we define its height as $h(\alpha)$. The height of a leaf node γ is set to zero, i.e., $h(\gamma) = 0$. For any node α, the height of its parent α^- is defined as one plus the height of α, that is, $h(\alpha^-) = h(\alpha) + 1$. The height of T', $h(T') = h(\lambda)$.

By applying Eq. (2) and Eq. (3), the SE of each node α is calculated, after which the computed SE_α values are appended to the set SEs (lines 15-16).

$$H_\alpha^{(1)} = -\sum_{i=1}^{|T_\alpha|} \frac{d_i}{vol(\lambda)} log \frac{d_i}{vol(\lambda)} \tag{2}$$

$$SE_\alpha = -\frac{g_\alpha}{vol(\lambda)} log \frac{vol(\alpha)}{vol(\lambda)} + H_\alpha^{(1)} \tag{3}$$

where $vol(\lambda)$ denotes the sum of degrees of all nodes within T_λ, g_α represents the total weight of the cut edges by T_α, and d_i indicates the degree of node i, where $i \in T_\alpha$.

Under a greedy strategy, 2D SE minimization iteratively merges the pair of nodes in T' that produces the largest $|\Delta SE|$, until no merge yields a negative ΔSE (lines 17–30), The calculation of ΔSE between two nodes is as follows:

$$\begin{aligned}
\Delta SE_{i,j} =& SE_{new} - SE_{old} \\
=& SE_{\alpha_n} - (SE_{\alpha_i} + SE_{\alpha_j}) \\
=& -\frac{g_{\alpha_n}}{vol(\lambda)} log \frac{vol(\alpha_n)}{vol(\lambda)} - \frac{vol(\alpha_i)}{vol(\lambda)} log \frac{vol(\alpha_i)}{vol(\alpha_n)} \\
& - \frac{vol(\alpha_j)}{vol(\lambda)} log \frac{vol(\alpha_j)}{vol(\alpha_n)} + \frac{g_{\alpha_i}}{vol(\lambda)} log \frac{vol(\alpha_i)}{vol(\lambda)} \\
& + \frac{g_{\alpha_j}}{vol(\lambda)} log \frac{vol(\alpha_j)}{vol(\lambda)}.
\end{aligned} \tag{4}$$

where α_i, α_j is non-root nodes in encoding tree T', the operation $MERGE(\alpha_i, \alpha_j)$ inserts a new node α_n into T' and remove α_i and α_j from T'. α_n satisfies: 1) its children consist of all the children of α_i and α_j; 2) its parent is the root node, i.e., $\alpha_n^- = \lambda$. The process is repeated until no more clusters can be merged. (lines 32-33). If no clusters can be merged in any subset, n is increased so that each subset contains more clusters, making further merges possible (lines 34–35). Finally, a complete encoding tree T of the integrated graph G' is generated, and the leaf nodes of each subtree of the root node in T form a sentence cluster. We denote the final clustering result as $\{C_1, C_2, .., C_c\}$.

2.3 Summary Extraction Based on Position-Aware Compression

To further extract summary sentences reflecting the core content of the input multi-document based on the above sentence clusters, we propose a position-aware compression method for summary extraction, as shown in Figure 2C.

Algorithm 1 Multi-relation Graph Clustering via 2D SE minimization

Input: Integrated graph $G' = (V, E)$, n: number of nodes in each subgraph.
Output: A partition P of V.

1: $P \leftarrow \{\{s\} \mid s \in V\}$
2: **while** $True$ **do**
3: $P_S \leftarrow$ take clusters from P in groups of size $\min(n, |\text{remaining clusters}|)$ without replacement
4: **for** each $P_s \in P_S$ **do**
5: $V' \leftarrow$ union of all the clusters in P_s
6: $E' \leftarrow$ get all edges relate to V' in E
7: $G'_s \leftarrow (V', E')$
8: SEs $\leftarrow \emptyset$
9: $T' \leftarrow$ add a root tree node λ
10: **for** each cluster $C \in P_s$ **do**
11: Add a node α to T', s.t. $\alpha^- = \lambda, T_\alpha = C$
12: **for** each sentence $s \in C$ **do**
13: Add a leaf node γ to T', s.t.
14: $\gamma^- = \alpha, T_\gamma = \{s\}$
15: Caculate SE_α via Eq. (3)
16: Append SE_α to SEs
17: **while** $True$ **do**
18: $P' \leftarrow \{\alpha \mid \alpha \in T,\ h(\alpha) = 1\}$
19: $\Delta SE \leftarrow \infty$
20: **for** each $\alpha_i \in P$ **do**
21: **for** each $\alpha_j \in P, j > i$ **do**
22: $\Delta SE_{ij} \leftarrow$ Eq. (4),
23: w/o actually merging α_i and α_j
24: **if** $\Delta SE_{ij} < \Delta SE$ **then**
25: $\Delta SE \leftarrow \Delta SE_{ij}$
26: $\alpha_{o_1} \leftarrow \alpha_i$
27: $\alpha_{o_2} \leftarrow \alpha_j$
28: **if** $\Delta SE < 0$ **then**
29: MERGE$(\alpha_{o_1}, \alpha_{o_2})$
30: **else**
31: **Break**
32: Append P' to P
33: **if** $|P_S| = 1$ **then**
34: **break**
35: **if** P unchanged since last iteration **then**
36: $n \leftarrow 2n$
37: **return** P

Specifically, for the j-th sentence s_{ij} in the i-th original document, its positional importance is defined as:

$$w_{pos}(s_{i,j}) = 1 - \frac{pos_{i,j} - 1}{N_i - 1}$$

where pos_{ij} represents the position index of the sentence in original document, and N_i denotes the total number of sentences in that document. Subsequently, we calculate the sum of w_{pos} for all sentences in each sentence cluster (e.g., C_i) to obtain the positional-aware importance $Score_{pos}(C_i)$ of

that cluster:

$$Score_{pos}(C_i) = \sum_{s \in Ci} w_{pos}(s)$$

Then, we rank all clusters based on this metric to obtain the sorted clusters $\{C'_1, C'_2, ..., C'_c\}$, and retain the top c^* clusters, namely $\{C'_1, C'_2, ..., C'_{c^*}\}$. Finally, following the approach of Boudin and Morin (2013), we use the K-shortest paths algorithm to find K shortest paths as compression candidates $\{su_k^{(i)}\}_{k=1}^{K}$ in cluster C'_i. We calculate scores based on keyphrases to rank the compression candidates, and concatenate the Top-1 compression candidate from each cluster in sequence to obtain the final summary.

3 Experimental Settings

3.1 Datasets

We conducted experiments on four datasets. DUC-2004 and Multi-News (Fabbri et al., 2019) are two commonly used news MDS datasets, primarily used to validate the effectiveness of MRGSEM-Sum. Additionally, to further validate the generalization ability of MRGSEM-Sum, we also conducted experiments on PubMed (Cohan et al., 2018) and WikiSum (Liu et al., 2018). PubMed is a dataset consisting of long documents, while WikiSum is a multi-document dataset from a non-news domain. Detailed statistics of these four datasets can be found in Table 1.

Table 1: Test dataset statistics.

Datasets	Test Dataset Size	Average Document Length	Docs Per Cluster	Average Summary Length
Multi-News	5622	2103.49	2.76	263.66
DUC-2004	50	5978.2	10.0	107.04
PubMed	5025	3224.5	1.0	209.5
WikiSum	2000	2238.2	40.0	101.2

3.2 Implementation Details

In the position-aware step, the top-C parameter is set to 13 for Multi-News, 4 for PubMed, and 12 for WikiSum, while all clusters are retained for DUC2004, and the subgraph size n is fixed to 3. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24GB memory. For evaluation, ROUGE scores are computed using ROUGE-1.5.5 via the `pyrouge`[2] toolkit, and BERTScore is calculated with the roberta-large model (Liu et al., 2019). For both metrics, we report the F1 scores. Since there is no randomness, we only report the results of a single run.

[2] https://github.com/andersjo/pyrouge

4 Experimental Results and Analysis

4.1 Comparative Experiments

We compared MRGSEM-Sum with a series of strong baseline methods. Specifically, we selected two graph-based unsupervised extractive methods, **LexRank** (Erkan and Radev, 2004) and **TextRank** (Mihalcea and Tarau, 2004), which capture key information in the document through graph models. In addition, we considered two similarity-based unsupervised methods, **MMR** (Carbonell and Goldstein, 1998) and **Centroid** (Radev et al., 2004), which focus on computing the similarity between text units (such as sentences) to extract summaries. To further enrich the comparison scope, we introduced several supervised strong baseline models, including **PG-Original** (See et al., 2017), **PG-MMR** (Lebanoff et al., 2018), **Copy-Transformer** (Gehrmann et al., 2018), and **Hi-MAP** (Fabbri et al., 2019). Furthermore, given the advantages of current large language models (LLMs) in text generation, we specifically selected **GPT-3.5-turbo-16k** and **GPT-4-32k**, two LLMs capable of handling long texts, to test our method against LLMs.

The experimental results are shown in Table 2 . We can observe that: (1) MRGSEM-Sum outperforms other unsupervised models. Specifically, on Multi-News, compared to the classic graph-based unsupervised extractive baseline model LexRank, MRGSEM-Sum achieved improvements of over 4.94, 1.96, 3.79 and 2.11 on R-1, R-2, R-SU and BS, respectively. There are also substantial improvements compared to the similarity-based unsupervised methods MMR and Centroid. Additionally, compared to the powerful graph clustering baseline model SummPip, MRGSEM-Sum also shows improvement, with increases of 0.89, 1.38, 0.79 and 0.33 on R-1, R-2 R-SU and BS, respectively, and our model does not require prespecifying the number of clusters in the extraction process. (2) Compared to supervised methods, MRGSEM-Sum also demonstrates strong competitiveness. Notably, on Multi-News, MRGSEM-Sum even outperforms the PG-Original and PG-MMR models, with an increase of 1.36 on R-1 compared to PG-Original, and an increase of 2.66 on R-1 compared to PG-MMR. These results indicate that

our model achieves performance comparable to supervised models without the need for large-scale training data or complex training processes. (3) On the BS metric, MRGSEM-Sum consistently achieves the best performance among all compared methods. (4) In the experiments with LLMs, we called the online API interfaces of these LLMs, and our standardized prompt was "Summarize the above article:". From the experimental results, it can be seen that the performance of MRGSEM-Sum on the Multi-News and DUC-2004 datasets is comparable to that of LLMs, while our model is more compact.

Table 2: The comparison between MRGSEM-Sum and baselines on news MDS datasets Multi-News and DUC-2004 (**Bold** values are optimal among all methods; underlined values are best among unsupervised methods).

Models	Multi-News				DUC-2004			
	R-1	R-2	R-SU	BS	R-1	R-2	R-SU	BS
Unsupervised								
LexRank	38.27	12.70	13.20	82.4	35.56	7.87	11.86	81.76
TextRank	38.44	13.10	13.50	81.98	33.16	6.13	10.16	81.64
MMR	38.77	11.98	12.91	83.31	30.14	4.55	8.16	80.54
Centroid	41.25	12.83	15.63	81.82	36.52	8.82	11.68	81.71
SummPip	42.32	13.28	16.20	84.18	36.30	8.47	11.55	81.79
MRGSEM-Sum	43.21	14.66	16.99	**84.51**	**49.23**	**15.36**	**23.47**	82.11
Supervised								
PG-Original	41.85	12.91	16.46	84.21	31.43	6.03	10.01	80.09
PG-MMR	40.55	12.36	15.87	81.93	36.42	9.36	13.23	78.01
Copy-Transformer	**43.57**	14.03	17.37	82.19	28.54	6.38	7.22	77.51
Hi-MAP	43.47	**14.89**	**17.41**	81.32	35.78	8.90	11.43	77.48
LLMs								
GPT-3.5-turbo-16k	37.78	9.77	12.98	83.53	35.27	10.22	11.60	85.19
GPT-4-32k	39.87	11.54	14.17	83.66	36.44	11.92	11.34	**85.79**

4.2 The Performance in Long Documents and Non-news Domain MDS Datasets.

In order to further validate the generalization ability of MRGSEM-Sum, we conducted comparative experiments on PubMed and WikiSum datasets with unsupervised methods LexRank, TextRank, Centroid, and SummPip, as well as LLMs GPT-3.5-turbo-16k and GPT-4-32k.

The experimental results are shown in Table 3. We can observe that: (1) MRGSEM-Sum consistently outperforms unsupervised baseline models on these two datasets, particularly demonstrating significant improvements on the PubMed dataset. Compared to the Centroid, MRGSEM-Sum achieved improvements of 11.34, 8.47, 8.1 and 2.06 on R-1, R-2, R-SU and BS, respectively. Even when compared with the powerful unsupervised baseline model SummPip, MRGSEM-Sum still achieved substantial performance gains. These results indicate that MRGSEM-Sum not only exhibits significant advantages on traditional news multi-document summarization datasets but also demonstrates strong adaptability and generalization ability on long document and non-news domain multi-document datasets like PubMed and WikiSum. (2) The performance of MRGSEM-Sum is comparable to LLMs and, in some cases, even superior. For instance, on the WikiSum dataset, our model showed an improvement of 2.52 on the R-1 compared to GPT-4-32k. This further confirms the advantages of MRGSEM-Sum over LLMs.

Table 3: The comparison between MRGSEM-Sum and baselines on PubMed and WikiSum (**Bold** values are optimal among all methods; underlined values are best among unsupervised methods).

Models	PubMed				WikiSum			
	R-1	R-2	R-SU	BS	R-1	R-2	R-SU	BS
Unsupervised								
LexRank	28.77	8.16	7.44	81.87	36.60	9.40	13.03	83.77
TextRank	27.73	7.47	6.88	81.69	36.89	8.79	12.98	84.17
MMR	28.31	8.54	7.40	81.70	32.91	6.85	9.98	83.77
Centroid	26.62	6.57	6.32	81.39	36.37	8.08	12.53	83.99
SummPip	29.67	12.63	7.46	82.87	37.84	9.95	13.11	84.13
MRGSEM-Sum	37.96	**15.04**	14.32	83.45	**39.07**	**10.91**	**14.06**	84.20
LLMs								
GPT-3.5-turbo-16k	39.30	11.60	**14.57**	84.34	32.15	6.46	9.69	86.87
GPT-4-32k	**41.78**	13.35	13.35	**84.18**	36.55	9.06	12.29	**87.15**

4.3 Ablation Experiments

To better understand the contributions of different modules to the model performance, we conducted ablation experiments on Multi-News and PubMed datasets. Firstly, we replaced the multi-relation graph with the approximate discourse graph (ADG) used in SummPip (i.e., w/o multi-relation graph). Secondly, we replaced the 2D SE with spectral clustering method and set the number of clusters according to the experimental configuration by (Zhao et al., 2020) (i.e., w/o 2D SE clustering). Lastly, we replaced the position-aware compression method with a more general compression method proposed by Boudin and Morin (2013) (i.e., w/o position-aware).

The results are shown in Table 4, it can be observed that when removing the multi-relation graph, 2D SE clustering, and position-aware modules, MRGSEM-Sum's R-1, R-2, R-SU and BS scores all exhibited varying degrees of decline. Removing the multi-relation graph and 2D SE clustering modules led to significant performance drops on both datasets. For instance, on the Multi-News dataset, removing the multi-relation graph module resulted in a 4.25 decrease in R-1, while removing the 2D SE clustering module led to a 5.35 decrease

in R-1. Additionally, we noticed that removing the position-aware module had a more significant impact on the PubMed dataset, which we speculate is due to PubMed being a dataset of long documents, making it more sensitive to positional information during the summarization process.

Table 4: Ablation experiments on Multi-News and PubMed.

Models	Multi-News				PubMed			
	R-1	R-2	R-SU	BS	R-1	R-2	R-SU	BS
MRGSEM-Sum	**43.21**	**14.66**	**16.99**	**84.51**	**37.96**	**15.04**	**14.32**	**83.45**
w/o multi-relation graph	38.96	11.86	13.54	83.94	30.77	11.35	9.60	79.92
w/o 2D SE clustering	37.86	11.65	13.05	83.86	34.94	14.12	11.45	82.23
w/o position-aware	42.06	14.12	15.69	84.43	23.37	11.97	4.20	83.07

4.4 The Impact of Position-aware Compression

In ablation experiment, we have validated the effectiveness of position-aware compression. Next, we designed an experiment to further explore the impact of position-aware compression. Specifically, for the clusters $\{C'_1, C'_2, ..., C'_{c^*}\}$ sorted and truncated based on position-aware importance, we selected $c^* = 7, 8, ..., 19$ as inputs to the method proposed by Boudin and Morin (2013), and observed the quality of the generated summaries. The results are presented in Table 5, where it can be observed that under the position-aware importance sorting, even when only a portion of clusters participated in the subsequent summary generation, it still yielded satisfactory results. Additionally, an increase in the number of clusters led to a decrease in summary quality. This once again underscores the necessity of position-aware compression.

Table 5: The Impact of Position-aware Compression

c^*	Multi-News			
	R-1	R-2	R-SU	BS
ALL	42.06	14.12	15.69	84.43
7	39.66	12.14	14.31	84.27
8	41.01	12.77	15.35	84.39
9	41.93	13.25	16.06	84.42
10	42.60	13.69	16.58	84.46
11	42.97	13.99	16.85	84.45
12	43.18	14.26	16.98	84.45
13	**43.26**	14.50	16.90	84.51
14	43.21	14.66	**16.99**	84.52
15	43.09	14.78	16.75	84.52
16	42.93	14.90	16.58	**84.55**
17	42.74	14.97	16.38	84.52
18	42.57	15.03	16.20	84.50
19	42.43	**15.06**	16.05	84.49

4.5 Runtime Analysis

To evaluate the complexity of the model, we conducted a runtime analysis. As shown in Table 6, the average runtime per sample of the MRGSEM-Sum method on both the PubMed and Multi-News datasets is lower than that of SummPip, indicating superior time efficiency.

Table 6: Runtime Analysis on PubMed and Multi-News datasets (average time per sample).

Method	Multi-News	PubMed
SummPip	40ms	98ms
MRGSEM-Sum	**37ms**	**41ms**

4.6 Human Evaluation

In addition to automatic evaluations, we also conducted human evaluations for MRGSEM-Sum. We recruited three graduate students with excellent English reading and writing skills as evaluators, who independently rated 100 randomly selected summaries generated by each model in the Multi-News test set, including TextRank, Centroid, SummPip, GPT-3.5-turbo-16k, GPT-4-32k, and MRGSEM-Sum. The evaluators were instructed to rate the generated summaries based on three criteria: Fluency, Consistency, and Coverage, using a 5-star scale where 1 star represents the worst performance and 5 stars the best, and the average score from the evaluators represents the final score.

Figure 3 displays the distribution of human evaluation scores for these six models. It can be observed that MRGSEM-Sum has a significantly higher proportion of 5-star ratings in Consistency and Coverage compared to the other methods, particularly LLMs. In terms of Fluency, the performance of MRGSEM-Sum is comparable to LLMs and significantly better than other unsupervised summarization models. These findings further validate the effectiveness of the proposed MRGSEM-Sum. Additionally, an interesting observation is that both the cluster-based unsupervised summarization model SummPip and MRGSEM-Sum exhibit excellent performance in the Coverage metric, indicating that the approach of clustering sentences first and then compressing them can significantly enhance the coverage of generated summaries.

4.7 Case Study

To provide a more intuitive demonstration of the performance of MRGSEM-Sum, we conducted a

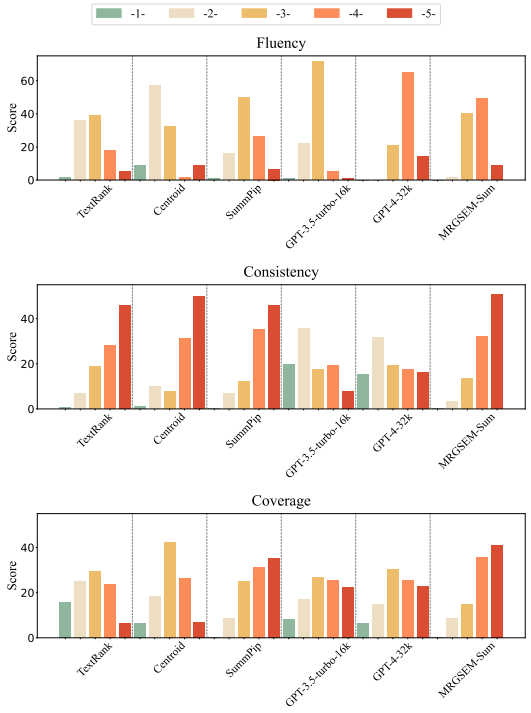

Figure 3: Human Evaluation.

Input : A 6.5 magnitude earthquake struck Eureka at around 4:30 p.m. Saturday , leaving much of the city without power and police scanners buzzing with reports of damage and injuries . The quake , which lasted for around 30 seconds , had power poles waving back and forth along Eureka streets and brought panicked residents out of their homes . The Bayshore Mall , as of 5:30 p.m. , had been closed off to customers and employees . It was unclear how many people were injured there , but one employee reported that several people were picked up by ambulances ...

Reference : – communities in humboldt county in northern california are cleaning up today after a 6.5-magnitude earthquake rocked the coastal area yesterday afternoon . " it was like the street was breakdancing , " a local resident who was driving when the temblor struck around 4 : 30pm told the times-standard of eureka , 33 miles from the epicenter . there were numerous reports of power outages and property damage , but no serious injuries and no tsunami warning . " it was a monstrous one , " a eurekar restaurateur told the los angeles times ...

LexRank: About 30 people visited hospitals for minor injuries , but therewere no reports of major injuries caused by the magnitude-6.5 temblor, which struck offshore at 4:27 p.m. about 33 miles southwest of thecoastal city of 26,000.More than 25,000 people were initially without power , but electricity was restored to everyone shortly after 6 a.m., according to David Eisenhauer , spokesman for Pacific Gas & ElectricCo ...

TextRank: At the 124-year - old St. Bernard 's Roman Catholic Church in Eureka , parishioners were busy picking up pieces of plaster that hadfallen from the ceiling and climbing ladders to 2) All along Broadway , Eureka 's busiest thoroughfare , lights wereout and traffic was heavy as people attempted to get out of town ...

Centroid: About 30 people visited hospitals for minor injuries , but therewere no reports of major injuries caused by the magnitude-6.5 temblor, which struck offshore at 4:27 p.m. about 33 miles southwest of thecoastal city of 26,000.More than 25,000 people were initially withoutpower , but electricity was restored to everyone shortly after 6 a.m., according to David Eisenhauer , spokesman for Pacific Gas & ElectricCo ...

SummPip: the media is slooooooooow says michael van poppel , who heads up the bno team twitter feed . arcata , near humboldt state university , had power , and residents of mckinleyville also reported reporting from los angeles and eureka , calif. - the day after a powerful earthquake rocked the northern california city of eureka , residents woke today to a mess : toppled chimneys , downed traffic signals and shattered nerves as minor aftershocks continued to rattle windows ...

GPT-3.5-turbo-16k: The articles describe a 6.5 magnitude earthquake that struck Eureka, California, causing significant disruption and damage. The earthquake occurred on a Saturday afternoon, affecting power lines, causing structural damage like fallen light fixtures and ceiling chunks at locations such as Bayshore Mall and Eureka Natural Foods ...

GPT-4-32k: The article reports on a 6.5 magnitude earthquake that hit Eureka at approximately 4:30 p.m. on a Saturday. The earthquake caused widespread power outages and significant damage, with reports of injuries. The quake lasted around 30 seconds and caused power poles to sway and residents to evacuate their homes in panic. Bayshore Mall was closed off to customers and employees by 5:30 p.m., and there were reports of injuries, although specific numbers were not yet clear ...

MRGSEM-Sum: a 6.5 magnitude earthquake struck eureka at around 4:30 p.m. saturday , leaving much of the city without power and police scanners buzzing with reports of damage and injuries . the quake , which lasted for around 30 seconds , had power poles waving back and forth along eureka streets and brought panicked residents out of their homes . the bayshore mall , as of 5:30 p.m. , had been closed off to customers and employees . employees also reported light fixtures falling out of the ceiling and floor tiles popping up off the ground . another employee reported seeing chunks of the ceiling fall onto customers ...

Figure 4: Case Study.

case study. We randomly selected an example from the results generated by the baseline model and MRGSEM-Sum on the Multi-News test set for qualitative analysis.

As shown in Figure 4, compared to other models, the summary generated by MRGSEM-Sum contains more detailed information and is closer to the reference summary. For example, the summary generated by MRGSEM-Sum includes key event elements, such as “A 6.5-magnitude earthquake struck eureka at around 4:30 p.m. saturday” and “without power”. These elements are highly consistent with both the input documents and the reference summary. Furthermore, MRGSEM-Sum also captures additional important details, such as “around 30 seconds” and “as of 5:30 p.m.”, which, although not in the reference summary, are present in the input documents and relatively significant. Even compared to LLs, the summary generated by MRGSEM-Sum is more informative. In contrast, unsupervised summarization models like SummPip miss key elements, such as “A 6.5-magnitude earthquake”, and extract irrelevant information not related to the news topic, such as “the media is slooooooooow”.

5 Conclusion

We present MRGSEM-Sum, an unsupervised multi-document summarization framework that addresses the dual challenges of complex relationships and information redundancy in MDS through three key innovations: (1) a multi-relational graph integrating both semantic and discourse relations to model sentence connections more comprehensively than single-relation approaches; (2) a 2D structural entropy minimization algorithm for adaptive, parameter-free clustering that automatically groups sentences while minimizing redundancy; and (3) a position-aware compression mechanism to distill salient information from clusters. Extensive experiments on Multi-News, DUC-2004, PubMed, and WikiSum demonstrate that MRGSEM-Sum outperforms unsupervised baselines and rivals supervised methods and LLMs in summary quality.

Limitations

Although our method demonstrates significant performance advantages in the unsupervised multi-document summarization task, its fluency is inferior to that of abstractive summarization due to its extractive nature, and the 2D structural entropy minimization algorithm relies on the quality of the graph construction.

Acknowledgments

References

Amanuel Alambo, Cori Lohstroh, Erik Madaus, Swati Padhee, Brandy Foster, Tanvi Banerjee, Krishnaprasad Thirunarayan, and Michael Raymer. 2020. Topic-centric unsupervised multi-document summarization of scientific and news articles. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 591–596. IEEE.

Nabil Alami, Mohammed Meknassi, Noureddine Ennahnahi, Yassine El Adlouni, and Ouafae Ammor. 2021. Unsupervised neural networks for automatic arabic text summarization using document clustering and topic modeling. *Expert Systems with Applications*, 172:114652.

Florian Boudin and Emmanuel Morin. 2013. Keyphrase extraction for n-best reranking in multi-sentence compression. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Janara Christensen, Stephen Soderland, Oren Etzioni, and 1 others. 2013. Towards coherent multi-document summarization. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1173.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of NAACL-HLT*, pages 615–621.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 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*, pages 1074–1084.

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

Salima Lamsiyah, Abdelkader El Mahdaouy, Bernard Espinasse, and Saïd El Alaoui Ouatik. 2021. An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. *Expert Systems with Applications*, 167:114152.

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.

Angsheng Li and Yicheng Pan. 2016. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory*, 62(6):3290–3339.

Miao Li, Jianzhong Qi, and Jey Han Lau. 2023. Compressed heterogeneous graph for abstractive multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13085–13093.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Prasenjeet Roy and Suman Kundu. 2023. Review on query-focused multi-document summarization (qmds) with comparative analysis. *ACM Computing Surveys*, 56(1):1–38.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1073–1083.

Nikolai Vogler, Songlin Li, Yujie Xu, Yujian Mi, and Taylor Berg-Kirkpatrick. 2022. An unsupervised masking objective for abstractive multi-document news summarization. *arXiv preprint arXiv:2201.02321*.

Jun Wang, Jinghua Tan, Hanlei Jin, and Shuo Qi. 2021. Unsupervised graph-clustering learning framework for financial news summarization. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 719–726. IEEE.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval*, pages 1949–1952.