Chain of Knowledge Graph: Information-Preserving Prompting for Noisy Multi-Document

* **Youngjin Lim**[1], * **Kangil Lee**[1], * **Jin-Woo Jang**[1], * **MinSu Shin**[1]
[1]LG Energy Solution, Seoul, Republic of Korea

Abstract

With the advent of large language models, the complexity of multi-document summarization task has been substantially reduced. The summarization process must effectively handle noisy documents that are irrelevant to the main topic while preserving essential information. Recently, Chain-of-Density (CoD) and Chain-of-Event (CoE) have proposed prompts to effectively handle the noisy documents by using entity-centric approaches for the summarization. However, CoD and CoE are prone to information loss during entity extraction due to their tendency to overly filter out entities perceived as less critical but that could still be important.

In this paper, we propose a novel instruction prompt termed as Chain of Knowledge Graph (CoKG) for multi-document summarization. Our prompt extracts entities and constructs relationships between entities to form a Knowledge Graph (KG). Next, the prompt enriches these relationships to recognize potentially important entities and assess the strength of each relation. If the acquired KG meets a predefined quality level, the KG is used to summarize the given documents. This process helps alleviate the information loss in multi-document summarization. Experimental results demonstrate that our prompt effectively preserves key entities and is robust to noisy documents.

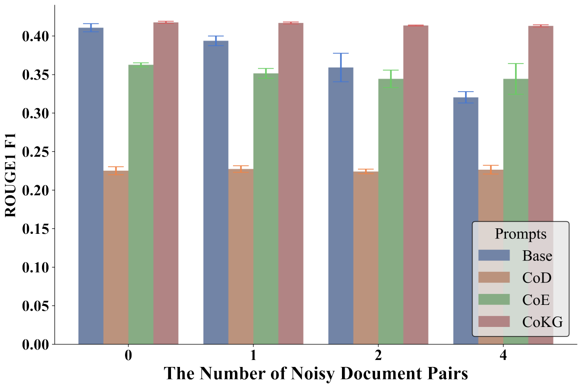

Figure 1: ROUGE-1 score charts showing the impact of adding noisy documents. Standard deviations are represented by error bars on each bar. 'Base' refers to a generic instruction such as "Summarize these documents". The low ROUGE1 scores of CoD and CoE indicate that these methods suffer from information loss. It is observed that the performance of 'Base' decrease as the number of noisy documents increase.

1 Introduction

The rise of foundation Large Language Models (LLMs) is redefining the landscape of various natural language processing (NLP) tasks. LLM-powered approaches are surpassing conventional supervised learning methods in tasks such as reasoning, sentiment analysis, and others, often with just a single prompt.

With this advancement, text summarization has entered a new phase (Pu et al., 2023). Instead of relying on 'golden' answers, texts can now be summarized with greater flexibility using LLMs. Furthermore, instructions enable precise tailoring of summary length and style. Examining prompt-based text summarization methods, we find that most of the recently proposed methods fall under the Chain-of-Thought (CoT) category (Zhang et al., 2024). Among the CoT approaches, CoD is a representative method (Adams et al., 2023). CoD starts with a sparse entity set and refines it iteratively to obtain a denser entity set while balancing detail and abstraction for summarization.

In real life, Multi-Document Summarization (MDS) is needed across various fields for diverse objectives. However, when collecting various documents from the web, such as news articles and community posts related to a specific event, it is common to come across documents that are irrelevant to the main topic. We define these documents as noisy documents. To effectively deal with MDS,

[*] All authors contributed equally to this work.

Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ COLING 2025, pages 1–5
January 20th, 2025. ©2025 NeusymBridge Workshop

CoE is recently proposed (Bao et al., 2024). CoE consists of four sequential steps: event extraction, event abstraction, statistical event analysis and final document summarization.

Although CoE and CoD prompts are designed to be robust to noisy documents, their prompts can lead to significant information loss in MDS 3. In CoE, relying solely on frequency to determine entity importance can result in the omission of contextually significant entities. Meanwhile, CoD can suffer from lack of entities due to stringent conditions to be entities.

For noise-robust and information-preserving summarization, we propose a novel instruction prompt termed as CoKG for MDS. CoKG aims to extract enriched entities to minimize information loss. Since our approach is entity-centric summarization, it is also robust to noisy documents. Our contributions are as follows.

- We propose a novel entity-centric MDS prompt that is relatively free from information loss by using knowledge graph.
- We demonstrate that chaining and expanding entities reduce information loss and enhance robustness to noisy documents.

2 Related Works

Text Summarization. Text summarization has two distinct tracks: extractive and abstractive summarization. In the context of neural machine approaches, extractive summarization is regarded as a combination of sequence labeling and selection tasks. (Nallapati et al., 2017; Zhou et al., 2018). Abstractive summarization is regarded as a sequence-to-sequence problem (See et al., 2017; Liu and Lapata, 2019) formulated as a source document $x = [x_1, ..., x_n]$ to a target summary $y = [y_1, ..., y_m]$, where n and m are the number of tokens.

Despite the effectiveness of supervised methods, scalability issue is still challenging. With the rise of LLMs, since LLMs can generate a summary with a few lines of instruction, prompt-based summarization has gained attention to address the issue (Kuznia et al., 2022; Liu et al., 2022; Adams et al., 2023).

MDS is similar to general text summarization but differs in that it integrates and deals with diverse perspectives on a single topic. For example, CoE minimizes irrelevant information and focuses on key events for a concise summary (Bao et al., 2024).

3 Chain of Knowledge Graph

3.1 Preliminary

Terminologies are defined as follows.

- $D = \{T_1, T_2 ..., T_N\}$ represents either a single document or set of documents.
- E_i represents a set of entities or events identified at iteration i.
- S_i represents a summary at iteration i.

For MDS, CoD starts by extracting an initial summary S_0 from D, with E_0 as an empty set. At each step $i \geq 1$, CoD identifies missing entities M_i by comparing D and previous summary S_{i-1} and following five conditions and six guidelines. These missing entities M_i are used to update E_{i-1}, resulting in E_i. Using E_i, CoD refines S_{i-1} into a new summary S_i. This process is repeated five times, resulting in a final summary after the last iteration. Meanwhile, CoE first extracts many events E_0 from D and consolidates E_0 into a set of key events A. Then, CoE identifies statistically the most common abstract events and utilizes the events to generate a summary S_0.

The stringent five conditions of CoD for adding entities limit the capacity to incorporate additional ones which leads to information loss during summarization. Thus, we propose a novel instruction prompt termed as CoKG.

3.2 Instruction Prompt

To address the information loss, our prompt constructs KG to preserve critical information as much as possible by enriching entities and relations. The constructed KG is effectively utilized to summarize the given documents.

CoKG prompt consists of six steps as follows (See Figure 2).

1. **Identify entities:** Identify and extract key entities E_i^o to minimize the influence of irrelevant information from D.
2. **Construct relations:** Construct relations between the elements of E_i^o. The relations are expressed as verbs, adjectives, and phrasal verbs.

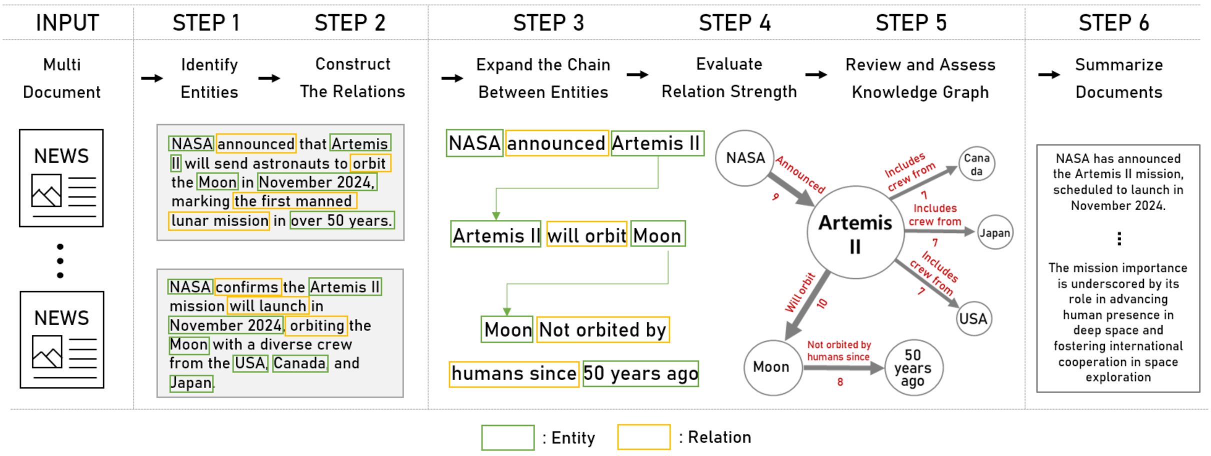
INPUT
STEP 1
STEP 2
STEP 3
STEP 4
STEP 5
STEP 6
Multi Document
Identify Entities
Construct The Relations
Expand the Chain Between Entities
Evaluate Relation Strength
Review and Assess Knowledge Graph
Summarize Documents
NEWS
NEWS
NASA announced that Artemis II will send astronauts to orbit the Moon in November 2024, marking the first manned lunar mission in over 50 years.
NASA confirms the Artemis II mission will launch in November 2024 orbiting the Moon with a diverse crew from the USA, Canada and Japan.
NASA announced Artemis II
Artemis II will orbit Moon
Moon Not orbited by
humans since 50 years ago
NASA
Announced
9
Artemis II
Includes crew from
7
Canada
Includes crew from
7
Japan
Includes crew from
7
USA
Will orbit
10
Moon
Not orbited by humans since
8
50 years ago
NASA has announced the Artemis II mission, scheduled to launch in November 2024.
⋮
The mission importance is underscored by its role in advancing human presence in deep space and fostering international cooperation in space exploration
: Entity
: Relation

Figure 2: Overall process to create a summary for multi-document. First, identify key entities and construct relationships between entities (STEP 1 and 2). Second, expand and chain the entities to create a knowledge graph, then evaluate the strength of the relationships and review the overall knowledge graph (STEP 3, 4, and 5). Third, generate a summary with knowledge graph (STEP 6).

3. **Expand the chain between entities:** Since each entity can have multiple relationships, it is necessary to sufficiently expand and chain entities ($E_i^o \rightarrow E_i$) to provide rich contextual information for summary.

4. **Evaluate relation strength:** Evaluate the entity connections by assigning scores ranging from 1 to 10, where 1 represents the weakest, 10 the strongest, and 5 a moderate link. This score depends on the strength of the relationships as found within the context of the documents.

5. **Review and assess knowledge graph:** Quantitatively evaluate the KG obtained from the previous four steps. If the KG effectively captures the main context and key entities of the given documents by including all relevant entities and relationships, assign a high score. The score ranges from 1 to 10, and if it does not achieve at least 7, return to Step 1 and reconstruct the KG.

6. **Summarize documents:** Finally, generate a summary based on the final KG and the given documents.

To prevent information loss, CoKG chains and expands entities. Furthermore, since CoKG is entity-centric summarization prompt, it is robust to noisy documents. Thus, the CoKG prompt can be considered an effective prompt for MDS.

Table 1: Evaluation results on the Multi-News dataset for assessing information loss. Even though CoKG compresses documents into key entity-focused graph, we find that CoKG experiences relatively no loss of information compared to the Base prompt.

	Base	CoD	CoE	CoKG
ROUGE-1	0.417	0.226	0.360	**0.418**
ROUGE-2	**0.114**	0.055	0.099	**0.114**
ROUGE-L	**0.189**	0.123	0.178	**0.189**
METEOR	0.266	0.111	0.195	**0.269**

4 Experiments

4.1 Datasets

To evaluate our instruction prompt for MDS, we use two datasets: Multi-News and PeerSum.

Multi-News consists of news articles and professional human-written summaries (Fabbri et al., 2019). Each summary includes links to the original articles cited. We use 100 randomly sampled sets of news collection from the test dataset.

PeerSum consists of review comments from OpenReview (Li et al., 2023). These comments range from official reviewers to public readers on a paper. The meta-review is considered as the reference summary. We used 100 randomly sampled sets of review collection from the test dataset.

4.2 Experimental Setup

Evaluation Metrics. We use the widely adopted ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) for evaluation. In addition, we uti-

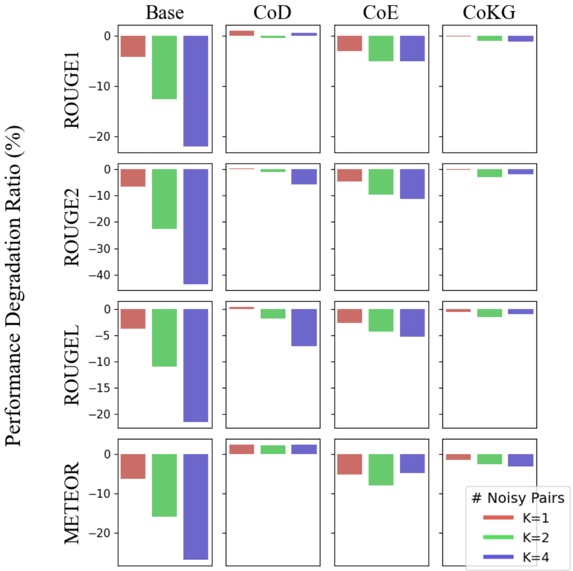

Figure 3: Performance degradation ratio resulting from the addition of noisy document pairs to the original set of documents. We use Multi-News dataset to obtain these results. K means the number of noisy document pairs. Performance degradation ratio represents the rate of performance drop when noisy documents are introduced.

Table 2: Evaluation results on the Multi-News and PeerSum datasets. We assess each prompt using G-Eval to evaluate summary quality from a human-friendly perspective. ↑ indicates that higher value is better, and ↓ indicates that lower value is better.

	Base	CoD	CoE	CoKG
Multi-News				
Coherence (↑)	4.301	3.990	4.467	**4.555**
Consistency (↑)	4.630	4.50	4.644	**4.685**
Fluency (↑)	2.731	2.587	2.680	**2.743**
Relevance (↑)	4.722	4.537	4.781	**4.818**
Average Rank (↓)	2.75	4	2.25	1
PeerSum				
Coherence (↑)	3.923	3.153	3.786	**4.090**
Consistency (↑)	**2.240**	1.833	2.047	2.107
Fluency (↑)	2.893	2.784	2.858	**2.946**
Relevance (↑)	**3.103**	2.412	2.898	3.059
Average Rank (↓)	1.5	4	3	1.5

lize G-eval (Liu et al., 2023) as a metric that has a high correlation with human evaluations for Natural Language Generation (NLG).

Comparison Prompts. To evaluate CoKG, we compare Base, CoD, and CoE prompts. The Base prompt is a generic instruction for summarization : *"Summarize the document below, which includes mutiple texts on similar topics."*

Model Selection. CoKG requires two abilities : decomposing instructions into several parts to easily handle each step and understanding the logical flow and connections. We selected Claude Sonnet 3.5 (ANTHROPIC, 2024) based on its state-of-the-art performance on decompositional and diagrammatic reasoning (Huang et al., 2024).

Noise Test. To evaluate robustness against noisy documents, we introduced text noise into a set of documents by appending unrelated article pairs both before and after the given document set.

4.3 Experimental Results

We evaluated how well the proposed prompt preserves information by recall-oriented metrics. Table 1 shows that CoKG exhibits almost no information loss compared to the Base prompt. In contrast, CoD and CoE show significant information loss. Table 2 demonstrates that the CoKG achieves competitive performance on two MDS datasets. In the Multi-News, our prompt achieves the best performance across all metrics. However, on the PeerSum, our approach performs relatively worse than the Base on two metrics. We infer that these results attribute to the entity expansion process.

Meanwhile, Figure 3 illustrates that CoKG is robust to noisy documents. As K increases, the performance drop for Base and CoE prompts becomes more pronounced, whereas CoKG shows relative robustness to noisy documents. Since CoD performs poorly when $K = 0$ compared to other prompts, it can be inferred that CoD does not perform well in MDS. Based on this, CoD may not be robust to noisy documents but rather unsuitable for MDS.

In conclusion, we find that CoKG effectively preserves information while also being robust to noisy documents.

5 Conclusion

MDS complexity is eased by advent of LLM-based approaches. However, the previous approaches often suffer from information loss. In addition, since generic prompt tends to show inferior performance to severely noisy document set, entity-centric approach is necessary.

Thus, we propose CoKG that is robust to noisy documents and has information-preserving property. CoKG maximally extracts topic-related entities to minimize information loss. In noise test, we observe that our approach is resilient to noise. In addition, the results from the Multi-News and

PeerSum benchmarks demonstrate that CoKG effectively preserves information and that its summaries closely align with human-generated ones. These findings suggest CoKG produces a reliable summary for multi-document.

References

Griffin Adams, Alexander Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. *arXiv preprint arXiv:2309.04269*.

ANTHROPIC. 2024. Claude 3.5 sonnet. Introducing Claude 3.5 Sonnet.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Songlin Bao, Tiantian Li, and Bin Cao. 2024. Chain-of-event prompting for multi-document summarization by large language models. *International Journal of Web Information Systems*, (ahead-of-print).

Alexander 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, Florence, Italy. Association for Computational Linguistics.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, et al. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *arXiv preprint arXiv:2406.12753*.

Kirby Kuznia, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Less is more: Summary of long instructions is better for program synthesis. *arXiv preprint arXiv:2203.08597*.

Miao Li, Eduard Hovy, and Jey Han Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. *arXiv preprint arXiv:2305.01498*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xiaochen Liu, Yang Gao, Yu Bai, Jiawei Li, Yinan Hu, Heyan Huang, and Boxing Chen. 2022. Psp: Pre-trained soft prompts for few-shot abstractive summarization. *arXiv preprint arXiv:2204.04413*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2024. A systematic survey of text summarization: From statistical methods to large language models. *arXiv preprint arXiv:2406.11289*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.