Considering Length Diversity in Retrieval-Augmented Summarization

Juseon-Do[1†], Jaesung Hwang[1†], [*]Jingun Kwon[1], Hidetaka Kamigaito[2], and Manabu Okumura[3]
[1]Chungnam National University, [2]Nara Institute of Science and Technology (NAIST)
[3]Institute of Science Tokyo
{doju00,hjs3545}@o.cnu.ac.kr
jingun.kwon@cnu.ac.kr
kamigaito.h@is.naist.jp
oku@pi.titech.ac.jp

Abstract

This study investigates retrieval-augmented summarization by specifically examining the impact of exemplar summary lengths under length constraints, not covered by previous work. We propose a Diverse Length-aware Maximal Marginal Relevance (DL-MMR) algorithm to better control summary lengths. This algorithm combines the query relevance with diverse target lengths in retrieval-augmented summarization. Unlike previous methods that necessitate exhaustive exemplar-exemplar relevance comparisons using MMR, DL-MMR considers the exemplar target length as well and avoids comparing exemplars to each other, thereby reducing computational cost and conserving memory during the construction of an exemplar pool. Experimental results showed the effectiveness of DL-MMR, which considers length diversity, compared to the original MMR algorithm. DL-MMR additionally showed the effectiveness in memory saving of 781,513 times and computational cost reduction of 500,092 times, while maintaining the same level of informativeness.

1 Introduction

Retrieval-augmented generation (RAG) is a promising approach in natural language processing (NLP) because it allows large language models (LLMs) to improve generation quality by leveraging a broader set of information from external resources via in-context learning (ICL) (Brown et al., 2020; Han et al., 2022; Guo et al., 2023; Izacard and Grave, 2021; Qiu et al., 2022; Su et al., 2022; Wang et al., 2023; Shao et al., 2023). Early efforts to retrieve exemplars have focused on a nearest neighbor (NN) method, that compares only query and exemplar relevance (Shin et al., 2021; Rubin et al., 2022). To further improve performance, exemplar-exemplar relevance comparisons or employing a two-stage approach for the retrieval have been studied (Ye et al., 2023; Guo et al., 2023; Ye and Durrett, 2023; Margatina et al., 2023).

However, despite the success of previous studies, the impact of summary lengths in the ICL for retrieval-augmented summarization has not been yet explored for better controlling summary lengths. Because better controlling summary lengths can improve summarization performance (Kwon et al., 2023; Miculicich et al., 2023), we propose to incorporate length diversity to construct a pool for the retrieval. We first conducted preliminary experiments to investigate how the exemplars' target summary lengths affect the summarization. Using advanced models such as ChatGPT (GPT-4-turbo-preview),[1] the generated summaries closely matched the retrieved target exemplar lengths, that implies that exemplar length information is crucial in retrieval-augmented summarization.

Our preliminary experiments led us to focus on diverse target length information in the retrieval from a pool of exemplars (§3.2). In this paper, we propose a Diverse Length-aware Maximal Marginal Relevance (DL-MMR) algorithm for retrieving exemplars by considering not only query relevance but also target length diversity. Unlike the previous MMR method (Carbonell and Goldstein, 1998), which computes scores for all pairs of exemplars to obtain relevance-based diverse exemplars, DL-MMR simplifies the process by storing only the target lengths. By skipping the scoring of all exemplar-exemplar pairs, DL-MMR additionally lowers computational cost and saves memory for building the pool of exemplars.

We conducted experiments on three sentence summarization benchmarks: the Google, BNC, and Broadcast datasets. Then, we performed an in-depth analysis to assess the effectiveness of our DL-MMR algorithm, demonstrating its robustness across the datasets with large target length gaps.

* corresponding author
† Equal Contribution

[1]https://chat.openai.com/

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2489–2500

Our DL-MMR significantly outperformed the NN method, that shows the effectiveness of considering length diversity. Furthermore, DL-MMR was comparable to the MMR retrieval method, while saving the memory of 781,513 times and the computational cost of 500,092 times without losing informativeness. Human evaluation results also showed that considering length diversity is effective for producing informative and concise summaries in retrieval-augmented summarization.[2]

2 Maximal Marginal Relevance

MMR. The NN-based exemplar retrieval approach considers only the relevance between the exemplars and query (Liu et al., 2022). Although this approach can retrieve the nearest neighbors of mostly similar exemplars, it may limit diversity. To address this issue, MMR selects exemplars that are relevant to the query while being diverse enough using the following equation (Ye et al., 2023):

$$\arg\max_{q_j \in D/T} (1-\lambda)\mathrm{Dist}(q, q_j) - \lambda \max_{q_i \in T} \mathrm{Dist}(q_j, q_i), \quad (1)$$

where λ is to control the balance between relevance and diversity, and Dist denotes similarity. Assuming a given query q and that we have already selected a set of $T = \{q_i\}$ exemplars, we select the next one using the Equation (1).

Diverse Length-aware MMR. Although better controlling summary lengths can improve summarization performance (Kwon et al., 2023; Miculicich et al., 2023), it has not been fully explored yet in retrieval-augmented summarization. Our preliminary experiments (in Sec. 3.2) demonstrated that generated summaries generally adhere to the retrieved target exemplar lengths, highlighting the importance of exemplar length information in retrieval-augmented summarization, because previous summarization methods have not assumed that the desired length is provided.

For this purpose, we propose the DL-MMR algorithm, that chooses exemplars from the exemplar pool, based on their similarity to a given query, while ensuring sufficient target length diversity among exemplars. Considering length diversity would prevent an LLM from adhering to a specific length. Algorithm 1 describes the process of choosing exemplars from the pool in the inference step by utilizing Equation (2) instead:

$$\arg\min_{q_j \in D/T} (1-\lambda)\mathrm{Dist}(q, q_j) - \lambda \min_{q_i \in T} \mathrm{Diff}(q_j, q_i), \quad (2)$$

[2]Our code is available at https://github.com/JuseonDo/DL-MMR.

Algorithm 1 Diverse Length-aware MMR

Input: exemplar pool $D = \{q_1 \dots q_n\}$, given test query q, the number of exemplar k, length difference $Diff$ and semantic distance $Dist$

Output: selected exemplars $T = \{q_1 \dots q_k\}$

$\mathbb{S} := [[Diff(q_i, q_j)]]_{q_i, q_j \in D}$ {pairwise length difference between exemplars in D}

$\mathbb{Q} := [Dist(q, q_i)]_{q_i \in D}$ {distance between query and exemplars in T}

$\mathbb{S}, \mathbb{Q} := Scale(\mathbb{S}), Scale(\mathbb{Q})$ {min-max scaling to transform values to be between 0 and 1}

$T := \{\}$

while $|T| < k$ **do**

  $\hat{q} :=$ Equation(2) {get the next exemplar based on Eq (2)}

  T.add($\hat{q}$)

end while

return T

where λ indicates a weight between relevance and length diversity. Diff represents the length difference. We use min-max scaling to convert values from Diff and Dist.

While MMR necessitates scoring all pairs of exemplars within the pool, resulting in a scoring count of $n(n-1)/2$, where n indicates the number of exemplars in the pool (Ye et al., 2023), DL-MMR calculates only the scoring count for the target length, which is n. Since the semantic similarity is a relative measure, we need to calculate all exemplar pair similarities for MMR. However, since the length information is a fixed value, we can immediately obtain it for DL-MMR. This additionally ensures significant memory and computational cost saving. However, please note both DL-MMR and MMR require recursive comparisons for exemplars in the inference step.

3 Experiments

3.1 Experimental Settings

Datasets. We used three sentence summarization benchmarks: Google (**Google**), Broadcast (**Broad**), and BNC (**BNC**) (Filippova and Altun, 2013; Clarke and Lapata, 2008). The Google dataset contains automatically created summaries based on the syntactic dependency trees from news headlines and the article's first sentence. The gold compression ratio for the test dataset is 0.45. The Broadcast and BNC datasets consist of human created summaries. The gold compression ratios for the test datasets are 0.76 and 0.72, respectively. Table 1 shows the dataset statistics.

Evaluation Metrics. The summary quality was evaluated using F_1 scores of ROUGE-1 (R-1), -2

Dataset	Training	Valid	Test	*Avg Src Len*	*Avg Tgt Len*
Google	200,000	1,000	1,000	24.4 ($\pm$9.2)	9.8 ($\pm$3.1)
Broad	-	-	1,370	19.8 ($\pm$12.8)	15.59 ($\pm$9.3)
BNC	-	-	1,629	27.9 ($\pm$15.3)	19.3 ($\pm$10.7)

Table 1: Statistics of datasets. The values in parentheses indicate the standard deviation of both the source and target lengths, respectively.

(R-2), and -L (R-L) (Lin, 2004), as well as the BERT score (BS) (Zhang* et al., 2020). To assess the summary length satisfiability, we calculated ΔCR, which is the difference between the model-generated and gold compression ratios (Kamigaito et al., 2018; Kamigaito and Okumura, 2020).

Implementation Details. We used `Llama2-13b -chat-hf` (Touvron et al., 2023), `Phi-3-Mini -128K-Instruct` (Abdin et al., 2024), and `GPT-4 -turbo-preview` (OpenAI et al., 2024) as our backbone. We used FAISS (Douze et al., 2024) to construct a pool and `bart-large` (Lewis et al., 2020) for measuring semantic distance. We used 8 exemplars and λ performed best in validation.

Compared Methods. The baseline retrieval methods were as follows: **Zero-shot** does not select exemplars from the pool; **Random** selects exemplars randomly from the pool; **NN** selects exemplars based on the nearest neighbor of the query using semantic similarity (Liu et al., 2022); **MMR** additionally incorporates relevance-based exemplar-exemplar diversity (Ye et al., 2023); and **DL-MMR** incorporates length diversity. We considered the length by either the compression ratio (DL-MMR$_{cr}$), the length in target word count (DL-MMR$_{tgt}$). Since the length in the source can offer diverse target lengths (Kwon et al., 2023), we also considered the source word count (DL-MMR$_{src}$). For both DL-MMR$_{tgt}$ and DL-MMR$_{cr}$, we used $\lambda = 0.1$. For DL-MMR$_{src}$, we used $\lambda = 0.5$. For MMR, we used $\lambda = 0.5$ on `Google`.[3]

3.2 Impact of Exemplar Lengths

We first examined how exemplar lengths affect retrieval-augmented summarization. We used `Google` as the dataset and tried to generate summaries by giving exemplars with a specific target compression ratio or word count. The exemplars with the desired target compression ratio or word count were randomly extracted from the pool. Table 2 shows the results. LLMs relied on the desired target compression ratio or word count in exemplars. These preliminary experiments led us to consider length diversity for retrieval-augmented summarization because typical summarization does not have specific target length information. Furthermore, both Llama-2-13b and GPT-4 faced difficulties when the exemplar lengths or ratios are large.

[3] Implementation details and validation performances on other datasets for λ are in Appendix A.

len	Llama-2-13b-chat-hf				GPT-4-turbo-preview			
	R-1	**R-2**	**R-L**	***gen***	**R-1**	**R-2**	**R-L**	***gen***
5	68.1	53.8	67.5	6.4	70.1	54.0	69.5	6.8
10	**76.1**	**64.3**	**75.2**	9.6	**75.5**	**63.5**	**74.7**	10.6
15	73.4	62.6	72.7	12.5	71.4	60.6	70.8	14.0
20	70.4	60.3	69.7	14.8	67.7	57.4	67.1	16.3
30%	74.6	62.2	73.9	37%	75.1	61.7	74.3	40%
50%	**75.8**	**64.0**	**74.9**	44%	**75.2**	**63.2**	**74.4**	48%
70%	73.1	62.0	72.3	54%	71.5	60.4	70.9	60%
90%	67.7	57.3	67.0	66%	66.3	56.1	65.7	74%

Table 2: Affect of exemplar lengths. *len* and *gen* indicate the desired length or ratio, and the generated length or ratio, respectively.

Data	Method	R-1	R-2	R-L	BS	Δ *CR*	*Cost*		
							Mem	*Time$_c$*	*Time$_i$*
Google	*Zero-Shot*	66.8	54.8	65.7	0.68	23.1	-	-	-
	Random	75.2	63.5	74.5	0.76	-3.8	-	-	0m02s
	NN	78.7	<u>67.9</u>	77.9	**0.79**	<u>-3.1</u>	-	-	17m58s
	MMR	78.9	68.7	78.2	**0.79**	-2.8	372G	11h06m	2h14m
	DL-MMR$_{cr}$	78.0	67.3	77.3	0.78	-1.5	3M	0m25s	17m58s
	DL-MMR$_{tgt}$	**79.1**	**69.0**†	**78.5**	**0.79**	**-0.7**†	476K	0m00s	17m58s
	DL-MMR$_{src}$	78.0	68.1	77.5	0.78	-1.0	588K	0m00s	17m58s
Broad	*NN*	<u>80.1</u>	<u>66.2</u>	<u>78.8</u>	0.77	<u>-4.5</u>	-	-	0m04s
	MMR	80.1	65.4	78.2	0.76	4.6	25M	0m28s	0m17s
	DL-MMR$_{cr}$	78.7	64.5	77.3	0.76	-6.5	28K	0m00s	0m04s
	DL-MMR$_{tgt}$	**81.9**†	**68.1**†	**80.7**†	**0.78**	**0.4**†	8K	0m00s	0m04s
	DL-MMR$_{src}$	81.5	67.6	80.4	0.78	-1.8	8K	0m00s	0m04s
BNC	*NN*	<u>74.5</u>	<u>58.8</u>	<u>72.1</u>	0.69	<u>-6.2</u>	-	-	0m03s
	MMR	75.8	59.7	73.0	0.70	-1.5	18M	0m22s	0m14s
	DL-MMR$_{cr}$	73.5	57.9	71.0	0.68	-8.9	20K	0m00s	0m03s
	DL-MMR$_{tgt}$	**76.6**†	**61.5**†	**74.3**†	**0.71**	**0.1**†	4K	0m00s	0m03s
	DL-MMR$_{src}$	76.0	60.8	73.6	0.70	-2.6	4K	0m00s	0m03s

Table 3: Experimental results using zero-shot, random, NN, MMR, and DL-MMR on Llama2-13b-chat-hf. *Mem* denotes the memory required to create the exemplar pool. *Time$_c$* and *Time$_i$* denote the time spent in constructing and loading exemplars in the inference step, respectively. † denotes the significant improvement (p<0.05) compared with NN. We used paired-bootstrap-resampling with 100,000 random samples (Koehn, 2004).

3.3 Retrieval-augmented Summarization

Table 3 shows the performance of Llama-2-13b-chat-hf on `Google`, `Broad`, and `BNC`. For `Google`, we used the `Google` training dataset as a pool. For `Broad` and `BNC` without their own training dataset, we used `BNC` and `Broad` datasets as a pool, respectively. DL-MMR significantly outperformed NN in R-2 and ΔCR. Considering length diversity for the retrieval improves ROUGE scores, though it

Method	R-1	R-2	R-L	BS	Δ *CR*
NN	76.0	<u>65.2</u>	75.5	0.75	<u>-4.7</u>
MMR	75.5	64.9	75.0	0.75	-4.8
DL-MMR$_{cr}$	75.3	64.6	74.8	0.74	**-2.6**
DL-MMR$_{tgt}$	**76.8**	**66.3**†	**76.3**	**0.76**	-2.7†
DL-MMR$_{src}$	74.2	63.1	73.5	0.73	-4.8

Table 4: Experimental results with Phi-3-mini-128k-instruct on `Google`. The notations are the same as those in Table 3.

	NN	MMR	DL-MMR$_{tgt}$	Gold
Conc.	<u>3.52</u>	3.59	**3.60**†	3.54
Infor.	3.54	3.51	3.57	**3.60**

Table 5: Human evaluation results. The notations are the same as those in Table 3.

does not always match the gold length. Utilizing the length in target word count outperformed the compression ratio and the length in source word count, which indicates the target length information is crucial in retrieval-augmented summarization. Furthermore, DL-MMR$_{tgt}$ was comparable to MMR while using 781,513 times less memory and being 500,092 times and 7 times faster than MMR in the construction and inference steps on `Google`, respectively. Table 4 shows the performance of the Phi-3-mini-128k-instruct model. DL-MMR$_{tgt}$ significantly outperformed both NN and MMR.

Source: Child mortality rates are dropping but are still high in some parts of the world.

Retrieved Exemplars.

1. **SRC w/ DL-MMR$_{tgt}$:** Some of the most vulnerable children are still waiting too long for adoption placements.
   TGT w/ DL-MMR$_{tgt}$: Some of the vulnerable children are still waiting too long for placements.
   SRC w/ MMR: Some of the most vulnerable children are still waiting too long for adoption placements.
   TGT w/ MMR: Some of the vulnerable children are still waiting too long for placements.
2. **SRC w/ DL-MMR$_{tgt}$:** Spanish fresh produce exports fell by four per cent year on year during the first quarter of 2009.
   TGT w/ DL-MMR$_{tgt}$: Spanish exports fell.
   SRC w/ MMR: Cholera is surging again in parts of the world, a World Health Organization expert said Thursday, pointing to epidemics in Nigeria and Cameroon.
   TGT w/ MMR: Cholera is surging in parts of the world.
3. **SRC w/ DL-MMR$_{tgt}$:** Children have gone missing from hospitals in Haiti raising fears of trafficking for adoption abroad.
   TGT w/ DL-MMR$_{tgt}$: Children have gone missing from hospitals in Haiti.
   SRC w/ MMR: New estimates show the US has the seventh highest cancer rate in the world.
   TGT w/ MMR: The US has the seventh highest cancer rate in the world.
4. **SRC w/ DL-MMR$_{tgt}$:** The World Bank has warned that world poverty is much greater than previously thought.
   TGT w/ DL-MMR$_{tgt}$: Poverty is greater than previously thought.
   SRC w/ MMR: Birth rates have dropped for a third year in a row in the United States.
   TGT w/ MMR: Birth rates have dropped for a third year in a row.
5. **SRC w/ DL-MMR$_{tgt}$:** The American Academy of Environmental Medicine has released its latest position paper on electromagnetic field and radiofrequency health effects calling for immediate caution regarding smart meter installations.
   TGT w/ DL-MMR$_{tgt}$: The American Academy of Environmental Medicine has released its paper on field and effects calling for immediate caution regarding smart meter installations.
   SRC w/ MMR: One in three children are now living in poverty and the figures are set to rise as budget cuts kick in, ministers were warned.
   TGT w/ MMR: One in three children are now living in poverty.
6. **SRC w/ DL-MMR$_{tgt}$:** British women are more likely to die in childbirth than those in the former communist state of Slovenia, new research has shown.
   TGT w/ DL-MMR$_{tgt}$: British women are more likely to die in childbirth than those in the former communist state.
   SRC w/ MMR: The rapid rise in child obesity may be levelling off, according to figures.
   TGT w/ MMR: The rise in child obesity may be levelling off.
7. **SRC w/ DL-MMR$_{tgt}$:** World Vision says as of today six million people are affected by new flooding in Pakistan.
   TGT w/ DL-MMR$_{tgt}$: Six million people are affected by new flooding in Pakistan.
   SRC w/ MMR: Streetism has become one of the major social problems facing humanity all over the world.
   TGT w/ MMR: Streetism has become one of the major social problems facing humanity.
8. **SRC w/ DL-MMR$_{tgt}$:** Birth rates have dropped for a third year in a row in the United States.
   TGT w/ DL-MMR$_{tgt}$: Birth rates have dropped for a third year in a row.
   SRC w/ MMR: A Congolese warlord has been jailed for 14 years by the International Criminal Court for using child soldiers.
   TGT w/ MMR: A warlord has been jailed for using child soldiers.

DL-MMR$_{tgt}$: Child mortality rates are dropping.
MMR: Child mortality rates are still high in some parts of the world.
Gold: Child mortality rates are dropping.

Table 6: Retrieved exemplars and output of Llama-2-13b-chat-hf from `Google`.

4 Analysis

Human Evaluation and Case Study. We sampled 100 sentences from `Google` for human evaluation. We assigned 40 evaluators, all of whom have obtained both a US high school and a US bachelor's degree, to rate the results from 1 to 5 (5 is the best) for conciseness (Conc) and informativeness (Infor). Table 5 shows the results. Considering diverse lengths is essential for producing concise and informative summaries.Table 6 shows the retrieved exemplars using DL-MMR$_{tgt}$ and MMR. It can retrieve exemplars with diverse target lengths.

Impact of Target Length Gaps. Since `Google` has a rather different compression ratio from `Broad` and `BNC` with similar compression ratios, we performed more experiments on `Broad` and `BNC` with the `Google` training dataset as a pool, to investigate the effect of large target length gaps. Table 7 shows the results. While retrieval with the use of DL-MMR$_{cr}$ and DL-MMR$_{tgt}$ is effective for summarization on both `Broad` and `BNC`, NN, DL-MMR$_{src}$, and MMR encounter difficulties with length generalization. This indicates the importance of considering length diversity for retrieval-augmented summarization. However, ΔCR was not sufficiently met even when using DL-MMR$_{tgt}$. Table 8 shows the results when we separated the test dataset into two, shorter or longer than the average target length (11 (Ghalandari et al., 2022)) in `Google`. Due to the relatively short summaries in `Google` used for the pool, even DL-MMR$_{tgt}$ encounters difficulties for relatively longer summaries. The results indicate that further improvements would be desirable for constructing a pool by considering target lengths in retrieval-augmented summarization.

Impact of Number of Exemplar. We conducted further experiments to better understand the impact of the number of exemplars on performance. Table 9 shows the results. We observed that at least four exemplars are required to improve performance while ensuring diversity in retrieval-

Data	Method	R-1	R-2	R-L	BS	Δ *CR*
Broad	*NN*	67.0	52.9	65.6	0.67	-22.8
	MMR	67.5	53.4	66.1	0.67	-23.8
	DL-MMR$_{cr}$	71.9	57.6	70.2	**0.71**	-16.51
	DL-MMR$_{tgt}$	**73.4**†	**60.1**†	**72.2**†	**0.71**	**-11.8**†
	DL-MMR$_{src}$	69.2	55.7	67.9	0.69	-19.0
BNC	*NN*	61.3	47.3	59.8	0.60	-27.1
	MMR	60.5	46.3	58.9	0.59	-27.2
	DL-MMR$_{cr}$	65.5	51.3	63.7	0.63	-19.7
	DL-MMR$_{tgt}$	**67.5**†	**53.6**†	**65.9**†	**0.64**	**-17.0**†
	DL-MMR$_{src}$	61.7	48.0	60.2	0.60	-23.5

Table 7: Experimental results with Llama-2-13b-chat on `Broad` and `BNC` using the `Google` as a pool.

Data	*tgt len*	R-1	R-2	R-L	BS	Δ *CR*	*cnt*
Broad	0∼11	**79.4**	**64.2**	**78.4**	**0.77**	**-0.6**	718
	12∼	66.8	55.5	65.4	0.66	-24.1	652
BNC	0∼11	**76.6**	**59.6**	**75.3**	**0.75**	**-0.5**	487
	12∼	63.6	51.1	61.9	0.61	-23.9	1,142

Table 8: Experimental results with Llama-2-13b-chat using DL-MMR$_{tgt}$. The *cnt* indicates the number of instances within each range.

augmented summarization.[4]

5 Related Work

Length Constraint. Text summarization has gained attention for controlling the output sequence length to produce concise summaries while preserving informativeness, because users often consider desired output lengths (Kikuchi et al., 2016; Takase and Okazaki, 2019; Kwon et al., 2023; Miculicich et al., 2023). Recently, LLMs have demonstrated remarkable zero-shot task-solving abilities, especially in instruction-based settings (Brown et al., 2020; Radford et al., 2019). Consequently, numerous studies have leveraged instruction-based approaches to control output sequence length, either by directly specifying the desired length (Juseon-Do et al., 2024), or by incorporating multiple control types such as constraints like greater or smaller than a given value (Jie et al., 2024).

Retrieval-Augmented Generation (RAG). RAG has been recognized as a promising method and has been investigated in various NLP tasks (Lee et al., 2019; Izacard and Grave, 2021; Rubin et al., 2022; Guo et al., 2023; Buettner and Kovashka, 2024). The core idea is to improve the quality of text generation by conditioning LLMs on carefully selected external exemplars. Preivous studies focused on retrieving the most relevant exemplars, which can cause bias, based solely on query–exemplar similarities (Rubin et al., 2022; Liu et al., 2022; Shin et al., 2021). Alternatively, a recent work considered exemplar-exemplar similarities with MMR (Goldstein and Carbonell, 1998) for a better chance to illustrate the required reasoning process (Ye et al., 2023).[5]

Num	Method	R-1	R-2	R-L	Δ *CR*
2	*NN*	75.5	63.3	74.6	-3.4
	MMR	75.3	62.9	74.5	-2.8
	DL-MMR$_{cr}$	71.6	59.6	70.8	4.0
	DL-MMR$_{tgt}$	73.2	62.2	72.4	7.4
	DL-MMR$_{src}$	**75.7**	**64.2**	**75.0**	**3.0**
4	*NN*	76.7	65.2	76.0	-3.6
	MMR	77.3	65.8	76.6	-3.5
	DL-MMR$_{cr}$	76.8	65.7	76.0	**-0.2**
	DL-MMR$_{tgt}$	76.2	65.5	75.4	2.8
	DL-MMR$_{src}$	**77.2**	**66.6**	**76.4**	0.6
6	*NN*	77.8	66.9	77.1	-3.2
	MMR	77.7	66.6	77.0	-3.0
	DL-MMR$_{cr}$	78.0	67.7	77.4	-1.2
	DL-MMR$_{tgt}$	**78.3**	68.0	**77.7**	-0.4
	DL-MMR$_{src}$	77.8	**68.0**	77.3	**-0.01**
8	*NN*	78.7	67.9	77.9	-3.1
	MMR	78.9	68.7	78.2	-2.8
	DL-MMR$_{cr}$	78.0	67.3	77.3	-1.5
	DL-MMR$_{tgt}$	**79.1**	**69.0**	**78.5**	**-0.7**
	DL-MMR$_{src}$	78.0	68.1	77.5	-1.0
10	*NN*	79.0	68.9	78.5	-2.9
	MMR	79.3	69.2	78.7	-2.9
	DL-MMR$_{cr}$	78.7	68.6	78.1	-2.0
	DL-MMR$_{tgt}$	**79.3**	**69.5**	**78.9**	**-0.8**
	DL-MMR$_{src}$	78.9	69.5	78.5	-1.4

Table 9: Experimental results of Llama-2-13b-chat-hf on `Google` in changing the number of exemplars.

6 Conclusion

We revealed that considering length diversity is crucial for retrieval-augmented summarization. To incorporate target length information, we proposed the DL-MMR algorithm, which allows us to obtain a wider range of exemplars with diverse lengths. Our analysis showed that DL-MMR outperforms MMR, resulting in memory and computational cost saving without losing informativeness.

[4] Additional experimental results are in Appendix B.

[5] Appendix C introduces other related work.

Limitations

While our DL-MMR was designed to better control summary lengths with reducing computational and memory costs, the performance gains might diminish as the number of exemplars decreases for obtaining length diversity. We conducted experiments, and the details are in Appendix D.

In addition, implementing DL-MMR may entail greater complexity than the NN method. To resolve this issue, we will release our code for future studies. Furthermore, while our DL-MMR works effectively in English, it might not be directly applicable to languages not covered by the exemplars in our database, especially those with different syntactic and morphological structures. We will extend our DL-MMR to multiple languages in the future to evaluate its robustness.

Acknowledgments

We would like to gratefully acknowledge the anonymous reviewers for their helpful comments and feedbacks. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155857, Artificial Intelligence Convergence Innovation Human Resources Development (Chungnam National University)).

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 481–490, Portland, Oregon, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Kyle Buettner and Adriana Kovashka. 2024. Quantifying the gaps between translation and native perception in training for multimodal, multilingual retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5863–5870, Miami, Florida, USA. Association for Computational Linguistics.

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*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

James Clarke and Mirella Lapata. 2006. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 377–384, Sydney, Australia. Association for Computational Linguistics.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression : an integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368, Lisbon, Portugal. Association for Computational Linguistics.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.

Demian Ghalandari, Chris Hokamp, and Georgiana Ifrim. 2022. Efficient unsupervised sentence compression by fine-tuning transformers with reinforcement learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1280, Dublin, Ireland. Association for Computational Linguistics.

Jade Goldstein and Jaime Carbonell. 1998. Summarization: (1) using MMR for diversity- based reranking and (2) evaluating summaries. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 181–195, Baltimore, Maryland, USA. Association for Computational Linguistics.

Zhicheng Guo, Sijie Cheng, Yile Wang, Peng Li, and Yang Liu. 2023. Prompt-guided retrieval augmentation for non-knowledge-intensive tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10896–10912, Toronto, Canada. Association for Computational Linguistics.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. 2022. Prototypical calibration for few-shot learning of language models.

Chiori Hori and Sadaoki Furui. 2004. Speech summarization: An approach through word extraction and a method for evaluation. *IEICE Transactions*, 87-D:15–25.

Tatsuya Ishigaki, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2019. Discourse-aware hierarchical attention network for extractive single-document summarization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 497–506, Varna, Bulgaria. INCOMA Ltd.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Prompt-based length controlled generation with multiple control types. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1067–1085, Bangkok, Thailand. Association for Computational Linguistics.

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *Sixth Applied Natural Language Processing Conference*, pages 310–315, Seattle, Washington, USA. Association for Computational Linguistics.

Juseon-Do, Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2024. InstructCMP: Length control in sentence compression through instruction-based large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8980–8996, Bangkok, Thailand. Association for Computational Linguistics.

Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2018. Higher-order syntactic attention network for longer sentence compression. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1716–1726, New Orleans, Louisiana. Association for Computational Linguistics.

Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2017. Supervised attention for sequence-to-sequence constituency parsing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 7–12, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hidetaka Kamigaito and Manabu Okumura. 2020. Syntactically look-ahead attention network for sentence compression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8050–8057.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, page 703–710. AAAI Press.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down rst parsing utilizing granularity levels in documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8099–8106.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2021. Improving neural RST parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1600–1612, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2023. Abstractive document summarization with summary-length prediction. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 618–624, Dubrovnik, Croatia. Association for Computational Linguistics.

Jingun Kwon, Naoki Kobayashi, Hidetaka Kamigaito, and Manabu Okumura. 2021. Considering nested tree structure in sentence extractive summarization with pre-trained transformer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4039–4044, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.

Lesly Miculicich, Yujia Xie, Song Wang, and Pengcheng He. 2023. Summarization with precise length control.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha, and Kristina Toutanova. 2022. Evaluating the impact of model scale for compositional generalization in semantic parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9157–9179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. page 9. OpenAI.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Selective annotation makes language models better few-shot learners.

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. 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 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Liangguo Wang, Jing Jiang, Hai Leong Chieu, Chen Hui Ong, Dandan Song, and Lejian Liao. 2017. Can syntax help? improving an LSTM-based sentence compression model for new domains. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1385–1393, Vancouver, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Xi Ye and Greg Durrett. 2023. Explanation selection using unlabeled data for chain-of-thought prompting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 619–637, Singapore. Association for Computational Linguistics.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Implementation Details and Hyperparameter Selection

Table 10 shows instructions for summarization.

For implementation details, we used NVIDIA RTX A6000. For the decoding step, we did set $\text{do}_{samples}$=False, length-penalty=1.0. The CPU used for calculations is an Intel 4th Gen Xeon Scalable Processor (16-core).

Table 11 shows the performance of the Llama-2-13b-chat-hf model on the `Google` validation dataset. Tables 12 and 13 show the performance of the Llama-2-13b-chat-hf model on `Broad` with using the `BNC` training dataset as the pool and `BNC` with using the `Broad` training dataset as the pool, respectively. We selected λ based on the best average ROUGE scores for each dataset.

Followings are the computational costs between MMR and our DL-MMR in Table 3.

Memory Usage.

- MMR Memory: 372 GB (or 372,000,000 KB)
- DL-MMR Memory: 476 KB
- Ratio: 781512.6

Time Spent to Score Similarities.

- MMR: 40007.4 seconds
- DL-MMR: 0.08 seconds
- Ratio: 500,092.5

Time Spent to Retrieve Exemplars in the Inference Step.

- MMR: 8090.6 seconds
- DL-MMR: 1077.7 seconds
- Ratio: 7.5

B Comparison to GPT-4

Table 14 shows the results on `Google` with using the `Google` training dataset as the pool. Our DL-MMR_{tgt} using Llama2-13b-chat-hf, which is relatively small, achieved comparable performance compared to NN_{gpt4}, which uses ChatGPT (GPT-4-turbo-preview). This indicates that considering diverse target length information is crucial for producing concise and informative summaries in retrieval-augmented summarization.

C Other Related Work

Sentence Compression. Sentence compression is the task of generating concise and informative summaries by removing unimportant words while preserving fluency. Following the success of tree trimming (Jing, 2000; Knight and Marcu, 2000; Hori and Furui, 2004; Clarke and Lapata, 2006; Berg-Kirkpatrick et al., 2011; Filippova and Altun, 2013), Filippova et al. (2015); Klerke et al. (2016); Wang et al. (2017) demonstrate the effectiveness of end-to-end neural network-based approaches. Kamigaito et al. (2018) introduce recursive attention modules that consider syntactic heads (Kamigaito et al., 2017), which can be extended to document-level summarization (Ishigaki et al., 2019), similar to graph neural networks in Kwon et al. (2021) leverage parsed discourse trees (Kobayashi et al., 2020, 2021). Kamigaito and Okumura (2020) demonstrate the effectiveness of syntactic recursive attention modules combined with the pre-trained language model BERT (Devlin et al., 2019). Reflecting the success of large language models (LLMs), Juseon-Do et al. (2024) highlight the usefulness of LLMs and their ability to control output length in sentence compression.

Task	Instruction
Sentence Summarization	Sentence:\n{src}\nSummary of the sentence without the less important words would be:\n

Table 10: Instruction format. The "src" indicates the placeholder for a source sentence.

Method	**R-1**	**R-2**	**R-L**	**BS**	**Δ *CR***
Zero-shot	68.7	57.2	67.9	0.69	22.4
Random	76.5	64.8	75.7	0.76	-4.4
NN	<u>78.8</u>	<u>68.0</u>	<u>78.1</u>	0.78	<u>-3.8</u>
MMR	79.4	68.8	78.7	0.78	-3.9
DL-MMR$_{cr}$	79.2	68.6	78.7	0.78	**-2.0**
DL-MMR$_{tgt}$	**79.9**†	**69.5**†	**79.2**†	**0.79**	-1.6†
DL-MMR$_{src}$	79.7	70.0	79.1	0.79	-1.5

Table 11: Experimental results of Llama-2-13b-chat-hf on the `Google` validation dataset with using the `Google` training dataset as the exemplar pool. The notations are the same as those in Table 3.

Method	λ	**R-1**	**R-2**	**R-L**	**Δ *CR***
NN	*0.0*	<u>80.1</u>	<u>66.2</u>	<u>78.8</u>	<u>-4.5</u>
MMR	*0.1*	79.8	65.2	78.2	-5.6
	0.2	79.8	65.6	78.3	-5.2
	0.3	79.9	65.9	78.4	-5.1
	0.4	79.4	65.4	77.9	-6.0
	0.5	79.4	64.9	77.7	-4.5
	0.6	79.3	64.9	77.5	-4.1
	0.7	79.4	64.9	77.6	-1.5
	0.8	79.4	64.9	77.5	1.3
	0.9	80.1	65.4	78.2	4.6
	1.0	80.1	65.4	78.2	5.5
DL-MMR$_{tgt}$	*0.1*	79.9	65.8	78.7	-4.5
	0.2	80.4	66.3	78.9	-3.9
	0.3	80.7	66.6	79.5	-2.8
	0.4	80.7	66.7	79.4	-3.1
	0.5	80.7	66.7	79.4	-1.3
	0.6	80.6	66.5	79.3	-1.8
	0.7	80.6	66.4	79.1	-0.8
	0.8	81.3	67.3	80.0	-1.8
	0.9	**81.9**†	**68.1**†	**80.7**†	**0.4**†
	1.0	81.5	67.6	80.2	-0.7

Table 12: Experimental results of Llama2-13b-chat-hf on `Broad` with using the `BNC` training dataset as the exemplar pool. The notations are the same as those in Table 3.

Method	λ	**R-1**	**R-2**	**R-L**	**Δ *CR***
NN	*0.0*	<u>74.5</u>	<u>58.8</u>	<u>72.1</u>	<u>-6.2</u>
MMR	*0.1*	74.6	58.7	72.1	-6.1
	0.2	75.0	59.1	72.6	-5.7
	0.3	74.6	58.9	72.1	-5.8
	0.4	75.1	59.4	72.6	-4.8
	0.5	75.4	59.7	72.9	-4.3
	0.6	75.1	59.5	72.6	-4.2
	0.7	75.4	59.3	72.7	-4.2
	0.8	74.8	58.6	72.0	-4.6
	0.9	75.4	59.4	72.6	-3.3
	1.0	75.8	59.7	73.0	-1.5
DL-MMR$_{tgt}$	*0.1*	74.7	58.9	72.3	-5.1
	0.2	75.5	60.0	73.2	-4.3
	0.3	75.5	60.1	73.2	-3.6
	0.4	75.7	60.6	73.5	-2.6
	0.5	76.0	60.9	73.7	-0.9
	0.6	**76.6**†	**61.5**†	**74.3**†	**0.1**†
	0.7	76.3	61.4	73.8	-0.2
	0.8	76.2	60.8	73.4	-0.2
	0.9	76.3	60.8	73.6	-0.4
	1.0	76.3	60.0	73.4	0.3

Table 13: Experimental results of Llama2-13b-chat-hf on `BNC` with using the `Broad` training dataset as the exemplar pool. The notations are the same as those in Table 3.

Method	**R-1**		**R-2**		**R-L**		**BERTScore**		**△ *CR***	
	valid	test	valid	test	valid	test	valid	test	valid	test
Zero-Shot	68.7	66.8	57.2	54.8	67.9	65.7	0.69	0.68	22.4	23.1
Random	76.5	75.2	64.8	63.5	75.7	74.5	0.76	0.76	-4.4	-3.8
NN	78.8	78.7	68.0	67.9	78.1	77.9	0.78	**0.79**	-3.8	-3.1
NN$_{gpt4}$	79.8	**79.1**	<u>68.2</u>	68.1	79.0	**78.5**	**0.79**	**0.79**	**-0.7**	**-0.3**
MMR	79.4	78.9	68.8	68.7	78.7	78.2	0.78	**0.79**	-3.9	-2.8
DL-MMR$_{cr}$	79.2	78.0	68.6	67.3	78.7	77.3	0.78	0.78	-2.0	-1.5
DL-MMR$_{tgt}$	**79.9**	**79.1**	69.5†	**69.0**	**79.2**	**78.5**	**0.79**	**0.79**	-1.6	-0.7
DL-MMR$_{src}$	79.7	78.0	**70.0**	68.1	79.1	77.5	**0.79**	0.78	-1.5	-1.0

Table 14: Experimental results based on zero-shot, random, NN, MMR, and DL-MMR based on Llama-2-13b-chat-hf and GPT-4-turbo-preview. † indicates the improvement is significant (p<0.05) compared with NN$_{gpt}$.