Sentence Extraction Framework with High Relevance and Divergence for Document Summarization

Huiwen Xue[1], Baoan Li[1], Denghao Ma[1,✉], Xueqiang Lv[1], and Xiaoxi Wang[2]

[1] Beijing Information Science and Technology University, Beijing, China
xue_huiwen@163.com, {liba,madenghao,lxq}@bistu.edu.cn
[2] China Agricultural University, Beijing, China wangxx0709@sina.com

Abstract. Document summarization task is to condense documents while retain their key information, playing a key role in processing large-scale textual data. However, large language models (LLMs) focus on modeling frequently occurring information but often overlook less frequent yet crucial details, leading to potential information loss. To address this limitation, we propose a novel solution, i.e., Sentence Extraction Framework with High Relevance and Divergence for Document Summarization (SERD). In the framework, *a two-channel document attention module* is designed for ensuring high relevance of the extracted sentences, and *a coarse-fine granularity synergy MMR module* is designed for ensuring high divergence. So SERD can capture low-frequency yet crucial information by balancing relevance and divergence. The extracted sentences of high quality are subsequently fed into generative models to generate summaries. SERD is a universal plug-in component for LLMs, overcoming their limitations. Extensive experiments on two public datasets verify that our solution outperforms the SOTA baselines and significantly enhances the summary generation capabilities of LLMs.

Keywords: Document Summarization · Sentence Extraction · Coarse-fine Granularity.

1 Introduction

With the rapid development of generative AI, the volume of documents produced by humans or AI is growing exponentially. As an opportunity, documents can provide us with valuable information. However, as a challenge, quickly and accurately understanding documents becomes a key obstacle in extracting information from them. To address this challenge, the task of document summarization has increasingly gained attention from both academic and industrial communities [16]. The task takes a document as input and outputs a shortened summary that accurately captures the key semantic information of the document.

The information "divergence" nature of the input document makes it challenging for Transformer-based models to capture all key information of the document,

✉ Corresponding author.

including pre-trained ones and LLMs. These models rely on the attention mechanism. They often prioritize frequently occurring information while overlook less frequent but crucial details [41], i.e., cannot model the divergent information, which can result in the loss of important information. Therefore, we use the extract-then-generate framework to supplement this information for enhancing the ability of generation.

First, considering *what the sentences can represent the divergent information of a document*, we use two quality requirements for extracted sentences. On the one hand, the extracted sentences should cover the document's semantics as precisely as possible. This leads to a quality requirement of high relevance. On the other hand, the extracted sentences are expected to cover the semantics of the document as completely as possible. To achieve the "complete" coverage, the extracted sentences should diverge semantically, for encompassing a broader range of the document's semantics. This leads to a quality requirement of high divergence.

Second, to achieve high relevance and divergence qualities, we propose a sentence extraction framework with high relevance and divergence for document summarization (SERD). The framework contains two modules, i.e., a *two-channel document attention module* for ensuring high relevance and a *coarse-fine granularity synergy MMR module* for ensuring high divergence. In the first module, the sentence-channel document attention module constructs the matchings between sentences and document by modeling the relations among all sentences in the original document. The keyword-channel document attention leverages the keywords to construct the matchings of sentences to document. In the second module, i.e., *coarse-fine granularity synergy MMR*, the deep matching results from the *two-channel document attention module* are taken as fine-grained similarity. Besides, a large language model (LLM) is applied to generate a pseudo-oracle, which is utilized to estimate coarse-grained similarity. Finally, both fine-grained and coarse-grained similarities are input into the *coarse-fine granularity synergy MMR module* to identify the sentences with high divergence.

The main contributions are presented as follows:
• We propose a new sentence extraction framework with high relevance and divergence to address the limitation of LLMs, which is a new universal plug-in component for LLMs.
• We propose a new solution to match sentences and a document for extracting sentences with high relevance, i.e., two-channel document attention module.
• We propose the coarse-fine granularity synergy MMR to ensure the high divergence of extracted sentences, which is an advancement in MMR.
• Experiments conducted on two datasets verify that our solution outperforms the SOTA baselines and enhances the summary generation capabilities of LLMs.

2 Related Work

2.1 Research on Summary Generation

• *Extractive summarization methods.* These methods form summaries by directly extracting key sentences or phrases from the original text, selecting only the most representative and informative parts through algorithms, such as LEAD [28], LexRank [5], and MatchSum [40].
• *Abstractive summarization methods.* Unlike extractive methods, abstractive summarization involves understanding the content of the original text and rephrasing its core information into new sentences, e.g., Highlight-Transformer [22], CTF-DPP [30], and GraphSum+RoBERTa [20].
• *Extract-then-generate summarization methods.* These solutions combine the advantages of both extractive and abstractive methods. Initially, this approach extracts key information from the original document, then generates a summary based on this information. For example, PG-BRNN [7] and Hi-MAP [6] utilize LSTM-based sentence selection modules followed by generation. The Hierarchical-Transformer [23] employs LSTM to predict all paragraphs, calculates their ROUGE-2 scores, and ranks them to select required paragraphs.

2.2 Research on Sentence Extraction

• *Graph-based methods.* TextRank [27] is a popular graph-based approach, where the text is constructed as a graph, each sentence represents a node, and the edges between nodes represent sentence similarity. This allows the algorithm to identify the most important sentences in the text. Similar to TextRank, LexRank is also a graph-based method, but it typically uses cosine similarity as the weight for edges between nodes to assess sentence similarity.
• *Cluster-based methods.* In these methods, sentences are treated as points in a multi-dimensional space, and clustering algorithms, such as K-means [26], are used for grouping them. The central sentences from each cluster can be selected as the important sentences. Hierarchical clustering [14], distinct from K-means' flat clustering, creates a tree-like hierarchical structure among sentences, allowing for more flexible determination of important sentences.
• *Frequency-based methods.* TF-IDF estimates the importance of sentences by the frequency of words in a document relative to distribution frequency on the entire corpus [33].
• Feature-based methods. In these methods, the importance of sentences is calculated based on a set of predefined features, such as length, position, frequency, similarity to title [3].
• *Deep learning-based methods.* Neural network-based approaches have made significant advancement in the field of sentence extraction. For instance, Recurrent Neural Networks (RNNs) [15], Long Short-Term Memory (LSTMs) [11], and Transformer models are used for comprehending the semantics of texts and extracting important sentences. Since Transformer's efficacy in avoiding long-term dependency issues and capturing global information, it has gained wider applications than RNN and LSTM [8, 29, 36].

Discussion. Our solution belongs to the category of *deep learning-based methods*. But different from existing *deep learning-based methods*, our solution explicitly models qualities of sentences from both relevance and divergence views. Specifically, we propose a new two-channel document attention module to deeply match sentences and documents, for modeling the relevance quality of extracted sentences. Additionally, we introduce a coarse-fine granularity synergy MMR to effectively model the divergence quality of extracted sentences. The extracted sentences of high relevance and divergence are applied to LLMs. LLMs obtain more divergent information of a document and thus address the limitations of LLMs. Our proposed framework is a universal plug-in component for LLMs, and can enhance their capabilities of summarization generation.

3 Methodology

3.1 Task Definition

By specifying the input and output, we can define the task of document summarization:
• Input: A document D contains a sequence of sentences $\{s_1, s_2, \cdots, s_n\}$. Each sentence s_i contains a sequence of words $\{w_1, w_2, \cdots, w_m\}$.
• Output: A summary text T=$\{w_1^*, w_2^*, \cdots, w_k^*\}$. T contains a sequence of words $\{w_1^*, w_2^*, \cdots, w_k^*\}$, where k is much less than $m \cdot n$. T is necessary to convey the key semantics of D.

3.2 Sentence Extraction Framework with High Relevance and Divergence

As motivated in Section 1 and 2, the two-stage framework of *extract-then-generate* is widely applied to systems of document summarization [31, 36], i.e., first extracting important sentences from document, and then generating an summary over the important sentences. In this paper, we focus on the research of sentence extraction to address the limitation of LLMs, i.e., LLMs tend to prioritize high-frequency information while overlook low-frequency but important information in a document.

First, we use the quality requirements (relevance and divergence) for sentence extraction. Second, we propose a Sentence Extraction Framework with High Relevance and Divergence (SERD) to extract sentences with high qualities. The framework of SERD is shown in Figure 1. The two-channel document attention model is proposed to deeply match sentences with the document for estimating the relevance of sentences. The coarse-fine granularity synergy MMR selects sentences with high divergence from the relevant sentences to ensure that the low-frequency but important sentences can be extracted. The extracted sentences with high relevance and divergence are subsequently fed into a generative model (e.g., Llama2, Llama3, ChatGLM2, or BART) to generate summaries.

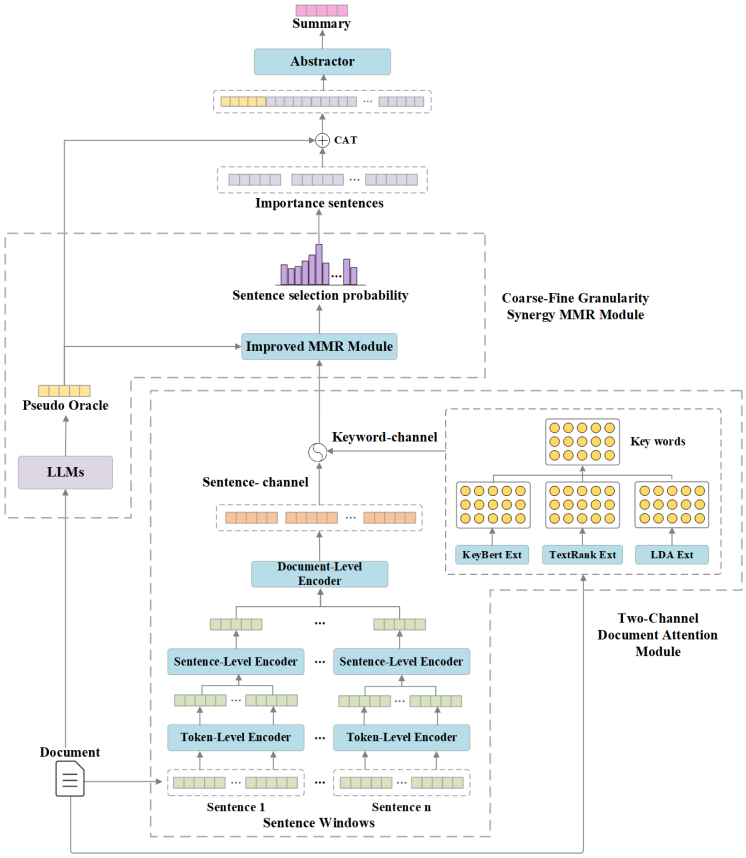

Fig. 1. Framework of SERD. The framework primarily consists of two-channel document attention module and coarse-fine granularity synergy MMR module.

Qualities of extracted sentences **Relevance.** A good summary should precisely capture the semantics of the original document [38]. In the two-stage framework, the precise summary generation hinges on the precise extraction of sentences. Therefore, the extracted sentences should capture the semantics of the document as precisely as possible, i.e., relevance.
Divergence. A good summary should as much encompass all the crucial points of the original document as possible. To this end, the extracted sentences are expected to cover the semantics of the document as completely as possible. To achieve "complete" coverage, the extracted sentences should exhibit semantic divergence, encompassing a broader range of document's semantics.

Two-channel document attention To ensure the high relevance of extracted sentences, we propose a new two-channel document attention module to deeply match each sentence $s_i \in D$ with the input document D. Due to the length of

D, encoding D directly using pre-trained language models is challenging, as they have a limitation on the number of words. This can result in a loss of information. To address this bottleneck, we propose the sentence-channel document attention model and the keyword-channel document attention model. In the first channel, we leverage the deep matchings between a sentence s_i and other sentences in D to replace direct matchings between s_i and D. This is because the number of sentences in D is much less than that of words in D. In the second channel, we extract keywords from D and use them to match s_i, as the number of extracted keywords is significantly lower than the total words in D. The two-channel document attention module aims to estimate the relevance of s_i to D from different perspectives, i.e., sentences and keywords.

Sentence-channel document attention model. In Figure 1, a document D is segmented into a sequence of sentences $S = \{s_1, s_2, \cdots, s_n\}$. Second, we encode each sentence s_i using a pre-trained language model, such as RoBERTa [24]. Third, we construct deep matchings among sentences by using self-attention mechanism [12]. During these matchings, the accurate representation and relevance of a sentence are learned by analyzing the relationships among sentences. We formulate the sentence-channel document attention model as follows:

$$e^*(s_i) = \sum_{s_j \in S} w(s_i, s_j) e(s_j), \tag{1}$$

$$r_s(s_i) = \sum_{s_j \in S} \frac{w(s_i, s_j)}{|S|}, \tag{2}$$

where $e(s_j)$ and $e^*(s_i)$ are the input and output embeddings of s_j. The $r_s(s_i)$ represents the relevance of sentence s_i to the document D. The $w(s_i, s_j)$ represents the relevance between the sentence s_i and s_j, and is estimated by Dot-Product.

Keyword-channel document attention model. In the field of summary generation, keywords of the input document have been verified important to enhance the summary generation process [9, 19]. Inspired by these solutions, we apply keywords to extract important sentences from the input document. But different from the above solutions, we design an evidence fusion method to extract keywords from the input document. Specifically, TextRank, LDA [1], and KeyBert [35] are performed to extract keywords. Each method extracts N keywords. All extracted keywords are ranked according to their frequencies, and the top N keywords are selected as the final set K.

We use the set K to estimate the relevance of s_i to D. Specifically, we use RoBERTa [24] to encode each keyword and obtain its embedding. Subsequently, we average the embeddings of all keywords, i.e., $e(K)$. Third, $e(K)$ is used for estimating the relevance of s_i, as follows:

$$r_{s,k}(s_i) = \phi_2(\omega_2[e^*(s_i); e(K); e^*(s_i) \otimes e(K); r_s(s_i)] + b_2), \tag{3}$$

where ϕ_2 with parameters ω_2 and b_2 is a $m \times n \times 1$ multi-layer perceptron with activation units ReLU, ReLU and Sigmoid. The $\otimes$ is the cross product. The $r_s(s_i)$ is the relevance of s_i to D, which is estimated by Equation 2.

Coarse-Fine Granularity Synergy MMR Maximal Marginal Relevance algorithm (MMR) [2] is a ranking algorithm used in information retrieval and natural language processing. We improve the MMR to ensure high divergence of extracted sentences. In existing text summarization solutions, some similarity estimation methods, such as cosine similarity and vector space model, have been applied to the MMR algorithm but cannot accurately measure the similarity between a sentence s_i and the input document D. Therefore, we propose a coarse-fine granularity synergy MMR to accurately estimate both fine-grained and coarse-grained similarities between s_i and D.

Intuitively, the deep matchings between s_i and D can be used for estimating the fine-grained similarity. So the relevance $r_{s,k}(s_i)$ from the two-channel document attention model can be taken as the fine-grained similarity. To estimate the coarse-grained similarity, we need to construct an embedding for D. Because D contains a large number of words and thus has extensive semantics, pre-trained language models cannot directly generate a good embedding for D. So we first use LLMs (e.g., Llama2) to generate a pseudo-oracle $\overline{D}$ for D, and then use a pre-trained language model (i.e., RoBERTa) to encode $\overline{D}$, denoted as $e(\overline{D})$. The cosine similarity of $e(s_i)$ and $e(\overline{D})$ is taken as the coarse-grained similarity. Therefore, we formulate the coarse-fine granularity synergy MMR model:

$$\begin{aligned} MMR+ = \arg\max_{s_i \in R \setminus S} \Big[& r_{s,k}(s_i) + \lambda \, Sim_1(e(s_i), e(\overline{D})) \\ & - (1-\lambda) \max_{s_j \in S} Sim_2(e(s_i), e(s_j)) \Big], \end{aligned} \tag{4}$$

where R represents the set of all candidates, S is the candidates that have already been selected, $R \setminus S$ represents the candidates not yet selected. The $r_{s,k}(s_i)$ is the deep semantic similarity of s_i to D, and estimated by using Equation 3. The $Sim_1()$ and $Sim_2()$ are cosine similarity functions.

Summary Generation Comparing with the traditional models, we use BART [18] as Abstractor. To validate that our solution can enhance the document summarization capabilities of LLMs, we use Llama2 [37], Llama3 [4], and ChatGLM2 [39] as Abstractor respectively. Since summary generation is not the focus of our contributions, we directly use Abstractor (e.g., Llama2, Llama3, ChatGLM2 or BART) to generate the final summary for D. The Abstractor is formulated:

$$Summary = Abs([e(S_n(D)); e(\overline{D})]), \tag{5}$$

where [] denotes the concatenate operation. The $S_n(D)$ is the top-n sentences extracted from D, and $e(S_n(D))$ is the embeddings of the sentences $S_n(D)$. The $\overline{D}$ is the pseudo-oracle of D, which is constructed and introduced in Section 3.2.

Model Optimization In the extracting stage, we use a greedy algorithm [36] to generate greedy summary sentence annotations. Greedy algorithm, in each iteration, individually combines a candidate sentence with previously selected

Table 1. Statistics of two datasets.

Datasets	Type	DocLen	SummaryLen	ExtNum
Multi-News	MDS	2103.49	263.66	10
Multi-XScience	MDS	778.08	116.44	10

Table 2. Example prompts for pseudo oracles generation. The output of this example shows that our prompt can effectively generate high-quality pseudo oracles.

Prompts: If you are a professional document summarizer, please provide a high-quality summary of the following multi-document. The content of the multi-document is as follows [input content].
Input content: In this Oct. 20, 2016, photo, a highway sign greets motorists heading into the small town of Whiteclay, Neb.The Lakota Hope faith ministry in Whiteclay has started a fundraising campaign to buy out the four beer stores that sell millions of cans annually in the tiny village next to the Pine Ridge Indian... Follow Grant Schulte on Twitter at https://twitter.com/GrantSchulte.The Solution Contrasted against the old liquor store buildings, the Makerspace will provide a safe place to work, access to art supplies, and state of the art equipment.The space will also provide an online store where our local artists can establish a more reliable income and teaching/mentoring opportunities.
Output: The Lakota Hope faith ministry in Whiteclay, Nebraska, has initiated a fundraising campaign to acquire and shut down four beer stores near the Pine Ridge Indian Reservation, which is grappling with high alcoholism rates. The ministry aims to raise at least $6.3 million to purchase the stores, which are located just 200 yards from the reservation and sold 3.5 million cans of beer in 2015 despite being in a town with only a dozen residents. The campaign, "B.O.B.S.Whiteclay P.L.A.N.," also seeks to promote the Lakotas as a nation. However, there are doubts about the effectiveness of the buyout, with some suggesting alternative uses for the spaces, such as a roller skating rink or park. The Nebraska Liquor Control Commission will meet to discuss the stores' licenses, amidst concerns about law enforcement in the village and the potential for increased intoxicated driving if the stores close. The ministry's efforts are met with both hope and skepticism from the community.

sentences and calculates ROUGE scores with the ground truth summary. The candidate sentence with the highest ROUGE score is selected as the annotated sentence in this iteration. During training of the extracting stage, our model first outputs the Top-K sentences, then optimized by cross-entropy loss between the Top-K sentences and the sentences selected by greedy method. In the generation stage, Abstractors generate a summary for a document. The cross-entropy loss between the generated summary and the ground truth summary is applied to optimize Abstractors.

4 Experiments

Datasets. The English datasets of Multi-News [6] and Multi-XScience [25] have been widely used in the task of document summarization [10, 36]. The two

datasets originate from the news and scientific literature domains. As part of our approach, we generate pseudo-oracles for these two datasets using LLMs (e.g., Llama2, Llama3, or ChatGLM2). Employing BART as the Abstractor, we use ChatGLM2 to generate pseudo-oracles. The lengths of the pseudo-oracles are determined by the average summary lengths observed in both datasets. We design prompts to guide the LLMs in generating pseudo-oracles that meet specific requirements. The example prompts are shown in Table 2. Statistics of datasets are shown in Table 1.

Comparison Solutions. For each dataset, we select its SOTA solutions and the results of LLMs as our baselines to test the effectiveness of our solution. The SOTA solutions on the dataset of Multi-News include Hierarchical-Transformer [23], PG-BRNN [7], Hi-MAP [6], CTF-DPP [30], GraphSum [20], GraphSum + RoBERTa [20], Highlight-Transformer [22], MatchSum [40], ChatGLM2 [39], Llama2 [37], and Llama3 [4]. The ROUGE-L score is not evaluated for MatchSum, as MatchSum utilizes summary-level ROUGE-L (i.e., ROUGE-LSum), which differs from the sentence-level ROUGE-L scoring criterion. The SOTA solutions on the dataset of Multi-XScience include LEAD [28], LexRanK [5], Hi-MAP [6], Pointer-Generator [34], BART-large [18], REFLECT [36], ChatGLM2, Llama2, and Llama3.

Evaluation Metric. ROUGE [21] is a widely-used metric for assessing natural language generation systems. It quantifies the similarity between the generated summaries and ground truth summaries. This paper assesses the quality of the generated summaries using ROUGE-1, ROUGE-2, and ROUGE-L metrics.

Implementation Details. In these experiments, the word-level encoding component of the multi-granularity hierarchical encoder is initialized with RoBERTa-base [24], while the sentence-level and document-level encoders are composed of a 3-layer Transformer. All our experiments were conducted in a hardware environment equipped with an NVIDIA A40 48GB, utilizing Python 3.9.17 and PyTorch 2.0.1 as the software framework. We use BART as Abstractor with an initial learning rate of 3e-5.

4.1 Experimental Results

Comparison of Overall Performance We evaluate all models on the two datasets of Multi-News and Multi-XScience, and report their results in Table 3 and 6, respectively. It can be seen that our solution significantly outperforms all SOTA baseline solutions and improves the document summary generation capability of LLMs. Highlight-Transformer uses an encoder-integrated highlighting mechanism to increase the attention weights of words in keywords. In the two-channel document attention model of our solution, the keywords are extracted and applied to estimate the relevance of sentences. The Hierarchical-Transformer model initially selects all sentences by using LSTM to estimate their ROUGE-2 scores. Our solution estimates the importance score of a sentence from both relevance and divergence perspectives. As a result, the extracted sentences effectively supplement less frequent yet important information. The use of LSTM

Table 3. Comparisons of experimental results on Multi-News corpus.

Model	R-1	R-2	R-L	Average	Average Imp
ChatGLM2-9B	33.45	10.73	19.15	21.11	+49.08%
Lead+BART	40.03	16.00	20.85	25.63	+22.79%
Highlight-Transformer	44.62	15.57	18.06	26.08	+20.67%
TextRank+BART	41.42	15.73	22.45	26.53	+18.62%
Hierarchical-Transformer	42.36	15.27	22.08	26.57	+18.44%
PG-BRNN	43.77	15.38	20.84	26.66	+18.04%
LexRank+BART	42.51	15.69	22.45	26.88	+17.08%
Hi-MAP	44.17	16.05	21.38	27.20	+15.70%
CTF-DPP	45.84	15.94	21.02	27.60	+14.02%
GraphSum	45.02	16.69	22.50	28.07	+12.11%
GraphSum+RoBERTa	45.87	17.56	23.39	28.94	+8.74%
MatchSum	46.20	16.51	-	-	-
SERD(BART)	**46.77**	**20.91**	**26.74**	**31.47**	-
Llama2-7B	32.34	10.20	28.42	23.65	+3.55%
SERD(Llama2-7B)	**33.62**	**10.69**	**29.15**	**24.49**	-
Llama3-8B	28.73	8.97	25.15	20.95	+13.57%
SERD(Llama3-8B)	**32.78**	**10.33**	**28.25**	**23.79**	-

Table 4. BERTScore and factual consistency on Multi-News corpus.

Model	BERTScore	Factual Consistency
BART-base	0.870	79.7
CTF-DPP	0.852	81.9
MatchSum	0.850	77.3
REFLECT	0.871	82.2
SERD	0.877	93.9

for encoding in the PG-BRNN and Hi-MAP models encounter long-term dependency and information loss issues. However, our solution utilizes the sentence-level attention mechanism in the sentence-channel document attention model, thus effectively addressing the issues of information loss. CTF-DPP improves the attention mechanism by using Determinantal Point Processes to calculate attention weights for generative summarization, thereby reducing information redundancy. Different from this, our solution extracts high-divergence sentences by using the coarse-fine granularity synergy MMR algorithm. Both GraphSum and GraphSum+RoBERTa utilize graph structures to explore paragraph relationships in encoders. In our solution, the sentence-channel document model leverages the attention technique to encode the relationships among sentences for capturing the deep matching between sentences and the input documents. The metric improvements achieved by our solution confirm its superior efficacy compared to baseline models and show that our solution effectively addresses the limitation of LLMs.

The work [36] applies BERTScore and factual consistency to evaluate the semantic similarity. The BERTScore highlights the similarity between the gen-

Table 5. Comparisons of divergence and relevance between MatchSum and SERD.

Model	Relevance	Divergence
MatchSum	22.51	18.71
SERD	22.75	19.91

Table 6. Comparisons of experimental results on Multi-XScience corpus.

Model	R-1	R-2	R-L	Average	Average Imp
Lead	27.46	4.57	18.82	16.95	+180.29%
LexRank	30.19	5.53	26.19	20.64	+130.18%
TextRank	31.51	5.83	26.58	21.31	+122.95%
Hi-MAP	31.66	5.91	28.43	22.00	+115.95%
Pointer-Generator	34.11	6.76	30.63	23.83	+99.37%
BART-large	33.29	8.07	17.31	19.56	+142.89%
REFLECT	34.18	8.20	17.42	19.93	+138.38%
ChatGLM2-9B	54.78	35.83	41.74	44.12	+7.68%
SERD(BART)	**57.72**	**39.47**	**45.34**	**47.51**	-
Llama2-7B	42.60	21.08	39.57	34.42	+2.06%
SERD(Llama2-7B)	**42.85**	**22.17**	**40.36**	**35.13**	-
Llama3-8B	43.01	21.42	40.43	34.95	+7.55%
SERD(Llama3-8B)	**46.30**	**23.50**	**42.98**	**37.59**	-

erated summary and the ground truth summary, and the factual consistency highlights the similarity between the generated summary and document. We also compare our solution to the better baselines on BERTScore and factual consistency and report the results in Table 4. Factual consistency is evaluated with FactCC. Besides, REFLECT [36] is a two-stage summarization generation model. The results show SERD achieves the best values of BERTScore and factual consistency, illustrating the generated summary by SERD is more similar to the ground truth summary in semantics.

Comparison of Relevance and Divergence We perform a group of experiments to verify the effectiveness of our solution on the relevance and divergence qualities, and the experimental results are presented in Table 5. Since MatchSum is the best baseline solution (see Table 3), we run our solution and MatchSum to show their relevance and divergence qualities. We use KL-divergence [17] to test the divergence quality among extracted sentences, and formulate it as $div(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$, where $div(P||Q)$ denotes the KL divergence from probability distribution P to Q. The $P(i)$ and $Q(i)$ respectively indicate the probabilities of word i in P and Q. P is the sequence of extracted sentences. Q is the sequence of ground truth summaries. The vector space model [32] is used for estimating the relevance quality of extracted sentences, i.e., $rel(\mathbf{P}, \mathbf{Q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$, where $\mathbf{P} = (p_1, \ldots, p_n)$ and $\mathbf{Q} = (q_1, \ldots, q_n)$, with n being the size of the vocabulary, and each p_i and q_i in the vectors representing the frequency of occurrence of the i-th word in the vocabulary within the

Table 7. Human evaluation. Ratings are based on a Liert scale of 1(worst) to 5(best). The "Flu","Inf", and "NR" represent Fluency, Informativeness, and Non-redundancy.

Model	Multi-XScience			Multi-News		
	Flu	Info	NR	Flu	Info	NR
ChatGLM2	3.088	3.260	2.816	3.312	3.624	2.704
SERD	3.640	3.472	3.620	3.748	3.480	3.756

Table 8. Results of ablation experiments.

	Ext			Abs	Performance				
row	SDAM	MMR+	KDAM	BART	R-1	R-2	R-L	Average	Avg Imp
r1	✓	✓	✓	✓	46.77	20.91	26.74	31.47	-
r2	✓		✓	✓	41.92	13.23	20.79	25.31	+24.34%
r3	✓	✓		✓	42.76	14.40	21.59	26.25	+19.89%
r4	✓			✓	39.35	12.81	20.53	24.23	+29.88%

respective paragraphs. In Table 5, it can be seen that the sentences extracted by SERD have higher relevance and divergence than those extracted by MatchSum.

Human Evaluation To evaluate the quality of the generated summaries, we conduct a human evaluation of the summaries generated by ChatGLM2 and our approach. ChatGLM2, being the best-performing LLM baseline (see Table 3 and 6), is selected for comparisons. Following the paper [13] that introduced a method of human evaluation, we focus on three critical aspects of summaries, i.e., *fluency*, *informativeness*, and *non-redundancy*. The fluency evaluates the structure and coherence of the summaries, the informativeness measures the amount of essential information captured, and the non-redundancy assesses the level of repetition within the summaries. For evaluation, we randomly select 50 samples from the Multi-News and Multi-XScience datasets, respectively. We enlist ten students as evaluators, each rating the summaries generated by different models for the same instance. Each instance receive ten evaluations, and the final score for each metric are calculated as the average of the ratings from all evaluators.

Results are presented in Table 7. We can see that our approach significantly outperforms ChatGLM2. **Fluency metrics** show that our approach achieves 3.64 and 3.748 scores on the two datasets, and ChatGLM2 achieves 3.088 and 3.312 scores. These comparisons illustrate a substantial improvement achieved by using our method. **Informativeness** shows that our approach outperforms ChatGLM2 on the Multi-XScience dataset and cannot achieve a significant improvement over ChatGLM2 on the Multi-News dataset. **Non-redundancy** shows that our approach achieves 3.62 and 3.756 scores on the two datasets, significantly higher than ChatGLM2 on both datasets. Overall, our approach outperforms ChatGLM2.

Table 9. The effectiveness of MMR+ and two-channel document attention model.

Model	R-1	R-2	R-L	Average
MMR	35.10	12.22	18.64	21.99
MMR+	36.36	13.40	18.93	22.90
RoBERTa	34.36	11.81	18.52	21.56
Two-Channel	34.92	12.19	18.90	22.00

Ablation Study To test the effectiveness of components in our proposed solution, we conduct a series of ablation experiments, and report the results in Table 8. SDAM denotes the sentence-channel document attention model, KDAM denotes the keyword-channel document attention model, MMR+ indicates the coarse-fine granularity synergy MMR model, and BART refers to Bidirectional and Auto-Regressive Transformers applied to our solution for generating the final summary. The ✓ denotes that the corresponding component is applied.

In Table 8, we can see that the metrics of row r1 are better than these of other rows. This is because SDAM, KDAM, and MMR+ are applied to our solution. When MMR+ is not applied, its metrics, i.e., the row r2, are much less than these of row r1. This verifies the effectiveness of MMR+. When KDAM is not applied, i.e., the row r3, the metrics of r3 are not as good as these of r1. This illustrates that KDAM is important to sentence extraction and also verifies the effectiveness of KDAM.

To further verify the effectiveness of MMR+, we run the original MMR model and show the results in Table 9. To clearly compare MMR+ and MMR, the generation model BART is not applied. According to the results, MMR+ significantly outperforms MMR, which illustrates MMR+ can select sentences with higher qualities than MMR. To further test the effectiveness of our two-channel document attention model, we use RoBERTa to replace the two-channel document attention model in SERD. The results are presented in Table 9 where the two-channel document attention model is denoted as Two-Channel. It can be seen that the two-channel document attention model performs better than RoBERTa, which verifies the effectiveness of the two-channel document attention model.

In the sentence-channel document attention model, multiple-layer Transformers are applied to model the relationships among sentences. We investigate the effect of layer number on the metrics by performing experiments and report the experimental results in Table 10. As shown in Table 10, the performance of the Transformer with N=3 layers surpasses the setups with N=2 and N=4 layers, leading to the selection of N=3 as the optimal configuration. In the MMR+ model, the λ is used for balancing multiple similarities. We investigate the effect of different λ values on metrics, and the results are presented in Figure 2. It can be seen that $\lambda = 0.5$ brings better metrics than other values.

Table 10. Impact of the number of Transformer layers.

N	ROUGE-1	ROUGE-2	ROUGE-L
2	32.22	10.22	17.71
3	33.65	11.64	18.18
4	33.01	11.16	17.89

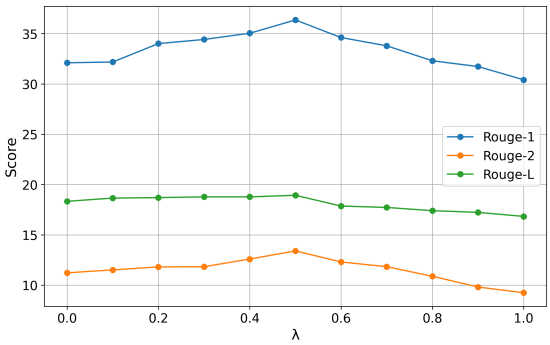

Fig. 2. The impact of λ on MMR+.

5 Conclusion

To study the document summarization task deeply, we observed that LLMs tend to prioritize high-frequency information while overlook low-frequency but important information when generating document summaries. To address this limitation, we applied two quality requirements (i.e., relevance and divergence) to extract sentences that contain important information. Besides, we propose a sentence extraction framework with high relevance and divergence (SERD) with a two-channel document attention model and coarse-fine granularity synergy MMR to capture and model the two qualities. Experimental results show that SERD achieves advanced performance on both the Multi-News and Multi-XScience datasets.

References

1. Blei, David M and et al.: Latent dirichlet allocation. Journal of machine Learning research, pp.993–1022 (2003)
2. Carbonell, Jaime and et al.: The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. SIGIR Forum, pp.209–210 (2017)
3. Christian, Hans and et al.: Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). ComTech: Computer, Mathematics and Engineering Applications, pp.285 (2016)
4. Dubey, Abhimanyu and et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783, (2024)
5. Günes Erkan and et al.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. J. Artif. Intell. Res., pp.457–479 (2004)

6. Fabbri, Alexander and et al.: Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. ACL, pp.1074–1084 (2019)
7. Gehrmann, Sebastian and et al.: Bottom-Up Abstractive Summarization. EMNLP, pp.4098–4109 (2018)
8. Guo, Mandy and et al.: Bottom-Up Abstractive Summarization. Findings of the NAACL, pp.724–736 (2022)
9. He, Junxian and et al.: CTRLsum: Towards Generic Controllable Text Summarization. EMNLP, pp.5879–5915 (2022)
10. Hewapathirana, Kushan and et al.: Multi-Document Summarization: A Comparative Evaluation. ICIIS, pp.19–24 (2023)
11. Hochreiter, Sepp and et al.: Long Short-Term Memory. Neural Computation, pp.1735-1780 (1997)
12. Zhongzhan Huang and et al.: Understanding Self-attention Mechanism via Dynamical System Perspective. ICCV, pp.1412-1422 (2023)
13. Jin, Hanqi and et al.: Abstractive multi-document summarization via joint learning with single-document summarization. EMNLP, pp.2545–2554 (2020)
14. Johnson, Stephen C.: Hierarchical clustering schemes. Psychometrika, pp.241–254 (1967)
15. Jordan, Michael I.: Serial order: A parallel distributed processing approach. Advances in psychology, 471–495 (1997)
16. Koh, Huan Yee and et al.: An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics. ACM Comput. Surv., pp.35 (2022)
17. Kullback, S. and et al.: On Information and Sufficiency. The Annals of Mathematical Statistics, pp.79–86 (1951)
18. Lewis, Mike and et al.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. ACL, pp.7871–7880 (2020)
19. Li, Haoran and et al.: Keywords-Guided Abstractive Sentence Summarization. Proceedings of the AAAI Conference on Artificial Intelligence, pp.8196-8203 (2020)
20. Li, Wei and et al.: Leveraging Graph to Improve Abstractive Multi-Document Summarization. ACL, pp.6232–6243 (2020)
21. Lin, Chin-Yew.: Rouge: A package for automatic evaluation of summaries. Text summarization branches out, pp.74–81 (2004)
22. Liu, Shuaiqi and et al.: Highlight-Transformer: Leveraging Key Phrase Aware Attention to Improve Abstractive Multi-Document Summarization. Findings of the ACL-IJCNLP, pp.5021–5027 (2021)
23. Liu, Yang and et al.: Hierarchical Transformers for Multi-Document Summarization. ACL, pp.5070–5081 (2019)
24. Yinhan Liu and et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692, (2020)
25. Lu, Yao and et al.: Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles. EMNLP, pp.8068–8074 (2020)
26. MacQueen, James and et al.: Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, pp.281–297 (1967)
27. Mihalcea, Rada and et al.: TextRank: Bringing Order into Text. EMNLP, pp.404–411 (2004)
28. Ani Nenkova and et al.: Automatic Summarization. Foundations and Trends® in Information Retrieval, pp.103-233 (2011)
29. Parnell, Jacob and et al.: A Multi-Document Coverage Reward for RELAXed Multi-Document Summarization. ACL, pp.5112–5128 (2022)

30. Perez-Beltrachini, Laura and et al.: Multi-document summarization with determinantal point process attention. Journal of Artificial Intelligence Research, pp.371–399 (2021)
31. Pilault, Jonathan and et al.: On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. EMNLP, pp.9308–9319 (2020)
32. Salton, G. and et al.: A vector space model for automatic indexing. Commun. ACM, pp.613–620 (1975)
33. Salton, Gerard and et al.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., pp.613–620 (1986)
34. See, Abigail and et al.: Get To The Point: Summarization with Pointer-Generator Networks. ACL, pp.1073–1083 (2017)
35. Prafull Sharma and et al.: Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling. Preprints, (2019)
36. Song, Yun-Zhu and et al.: Improving Multi-Document Summarization through Referenced Flexible Extraction with Credit-Awareness. NAACL, pp.1667–1681 (2022)
37. Touvron and et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, (2023)
38. Jin-ge Yao and et al.: Recent advances in document summarization. Knowl. Inf. Syst., pp.297–336 (2017)
39. Aohan Zeng and et al.: GLM-130B: An Open Bilingual Pre-trained Model. ICLR, (2023)
40. Zhong, Ming and et al.: Extractive Summarization as Text Matching. ACL, pp.6197–6208 (2020)
41. Krzysztof C and et al.: Rethinking attention with performers. ICLR, (2021)