Unsupervised Multi-document Summarization with Holistic Inference

Haopeng Zhang[1] **Sangwoo Cho**[2] **Kaiqiang Song**[2]
Xiaoyang Wang[2] **Hongwei Wang**[2] **Jiawei Zhang**[1] **Dong Yu**[2]
{haopeng,jiawei}@ifmlab.org
{riversong,swcho,shawnxywang,hongweiw,dyu}@global.tencent.com
[1]IFM lab, University of California, Davis [2] Tencent AI Lab, Bellevue, WA

Abstract

Multi-document summarization aims to obtain core information from a collection of documents written on the same topic. This paper proposes a new holistic framework for unsupervised multi-document extractive summarization. Our method incorporates the holistic beam search inference method associated with the holistic measurements, named Subset Representative Index (SRI). SRI balances the importance and diversity of a subset of sentences from the source documents and can be calculated in unsupervised and adaptive manners. To demonstrate the effectiveness of our method, we conduct extensive experiments on both small and large-scale multi-document summarization datasets under both unsupervised and adaptive settings. The proposed method outperforms strong baselines by a significant margin, as indicated by the resulting ROUGE scores and diversity measures. Our findings also suggest that diversity is essential for improving multi-document summary performance.

1 Introduction

The multi-document summarization (**MDS**) is one of the essential tools to obtain core information from a collection of documents written for the same topic. It seeks to find the main ideas from multiple sources with diversified messages. In spite of recent advances in MDS system designs (Mihalcea and Tarau, 2004; Liu and Lapata, 2019a; Xiao et al., 2022), three major challenges hinder its development:

First, existing extractive multi-document summarization systems rely on optimization with *individual* scoring. It becomes sub-optimal when we need to extract multiple summary sentences (Zhong et al., 2020). A typical individual system scores each candidate summary with only measurements of the newly added sentences during inference.

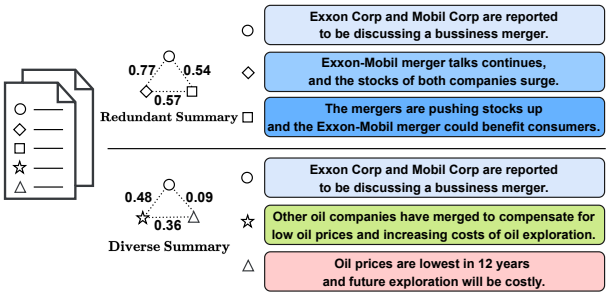

Figure 1: An example of a diverse summary vs. a redundant summary. Sentences in the redundant summary have higher semantic similarity than a diverse summary.

In contrast, the holistic system simultaneously measures all summary sentences and the relations among them. Despite recent efforts in holistic methods on a single document (An et al., 2022; Zhong et al., 2020), how to extract sentences holistically for multi-document summarization remains open. In this work, we propose an inference method that holistically optimizes the extractive summary under multi-document setting.

Second, multi-document summarization naturally contains excessively redundant information (Lebanoff et al., 2018). An ideal summary should provide important information with diversified perspectives (Nenkova and McKeown, 2011). In Figure 1, we show a salient and diversified summary versus a salient but redundant summary. A salient and diversified summary often covers the information thoroughly, while the salient but redundant summary is usually incomplete. Different from existing approaches (Suzuki and Nagata, 2017; Cho et al., 2019b; Xiao and Carenini, 2020) for limiting the repetitions, we introduce **S**ubset **R**epresentative **I**ndex (**SRI**), a holistically balanced measurement between importance and diversity for extractive multi-document summarization.

Finally, recent deep learning-based supervised summarization methods are data-driven and require a massive number of high-quality summaries in the

*Work done during Haopeng Zhang's internship at Tencent AI Lab Seattle.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 123–133
November 1–4, 2023.

training data. Nevertheless, hiring humans to write summaries is always expensive, time-consuming, and thus hard to scale up. This problem becomes more severe for multi-document summarization, since it requires more effort to read more documents. Therefore, existing multi-document summarization datasets are either small-scale (Over and Yen, 2004; Dang and Owczarzak, 2008) or created by acquiring data from the Internet with automatic alignments (Fabbri et al., 2019; Antognini and Faltings, 2020) that could be erroneous. Here we propose an unsupervised multi-document summarization method to tackle the low-resource issue. It can further benefit the unsupervised multi-document summarization, with the adaptive setting using large-scale high-quality single-document summarization data (e.g., CNN/DailyMail (Hermann et al., 2015)).

In this work, we present a novel framework for unsupervised extractive multi-document summarization, aiming to holistically select the extractive summary sentences. The framework contains the holistic beam search inference method associated with the holistic measurements named **SRI** (**S**ubset **R**epresentative **I**ndex). The SRI is designed as a holistic measurement for balancing the importance of individual sentences and the diversity among sentences within a set. To address data sparsity, we propose to calculate SRI in both unsupervised and adaptive manners. Unsupervised SRI relies on the centrality from graph-based methods (Erkan and Radev, 2004; Mihalcea and Tarau, 2004) for subset importance measurement, while adaptive SRI uses BERT (Devlin et al., 2018) fine-tuned on single document summarization (SDS) corpus for sentence importance measurement. Our method shows performance improvements in both the summary informativeness and diversity scores, indicating our approach can achieve better coverage of documents while maintaining the gist information of multi-documents. We highlight the contributions of our work as follows:

- We propose a novel holistic framework for multi-document extractive summarization. Our framework incorporates a holistic inference method for summary sentence extraction and holistic measurement called Subset Representative Index (SRI) for balancing the importance and diversity of a subset of sentences.
- We propose two unsupervised ways to measure SRI by using graph-based centrality or adapting from a single document corpus.
- We conduct extensive experiments on several benchmark datasets, and the results demonstrate the effectiveness of our paradigm under both unsupervised and adaptive settings. Our findings suggest that effectively modeling sentence importance and pairwise sentence similarity is crucial for extracting diverse summaries and improving summarization performance.

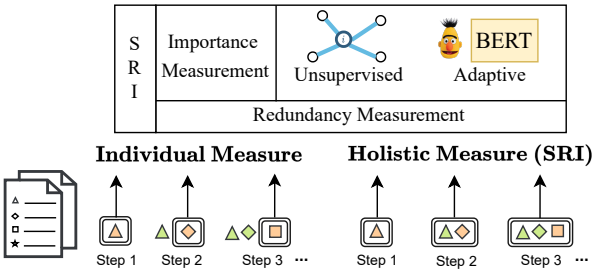

Figure 2: Illustration of the proposed holistic framework for multi-document summarization. The individual inference only resorts to each candidate while the holistic inference is based on all candidates. Orange and Green indicate newly added sentences and already added ones to the summary respectively.

2 Related Works

Multi-document Summarization Traditional non-neural approaches to multi-document summarization have been both extractive (Carbonell and Goldstein, 1998; Erkan and Radev, 2004; Mihalcea and Tarau, 2004) and abstractive (Ganesan et al., 2010). Recent neural MDS systems rely on Transformer-based encoder-decoder model to process the integrated long documents with hierarchical inter-paragraph attention (Liu and Lapata, 2019a; Fabbri et al., 2019), or attention across representations of different granularity (Jin et al., 2020). This work focuses on unsupervised MDS scenarios where gold reference summaries are unavailable. Prior unsupervised MDS systems are mostly graph-based (Erkan and Radev, 2004; Liu et al., 2021). Similar to our adaptive setting, Lebanoff et al. (2018) proposed to adapt the encoder-decoder framework from a single document corpus, but our work focuses on extractive summarization setting with holistic inference.

Sentence Importance Measurements Most works formulate extractive summarization as a sequence classification problem and use sequen-

tial neural models with different encoders like recurrent neural networks (Cheng and Lapata, 2016; Nallapati et al., 2016) and pre-trained language models (Liu and Lapata, 2019b; Zhang et al., 2023b). The prediction probabilities are treated as the importance measurement of sentences. On the other hand, unsupervised graph-based methods calculate the importance of sentences with node centrality and rank them for the summaries, including TextRank (Mihalcea and Tarau, 2004), LexRank (Erkan and Radev, 2004), PACSUM (Zheng and Lapata, 2019), and its variants (Liang et al., 2021; Liu et al., 2021). Recent researches (Xu et al., 2019; Wang et al., 2020; Zhang et al., 2022, 2023a) have explored Graph Neural Networks to obtain better representations for each sentence. Graph methods have merits in considering implicit document structure and to adapt with regardless of the input length.

graph neural network message passing or simply

Redundancy Considering only the importance of sentences for the summary leads to repeated information, and resolving the redundant contents is an essential problem in the extractive summarization system. Traditional methods to tackle redundancy relies on discrete optimization problem like Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998), Determinantal Point Process (DPP) (Kulesza and Taskar, 2012), and submodular selection (Lin et al., 2009). Trigram blocking is introduced to explicitly reduce redundancy by avoiding sentences that share a 3-gram with the previously added one (Liu and Lapata, 2019b). Paulus et al. (2017) first adopt trigram blocking in decoding for abstractive summarization. Ma et al. (2016) proposed the sentence filtering and beam search methods to for extractive summarization sentence selection. Xiao and Carenini (2020) conducted a systematic study of redundancy in long documents.

3 Method

This section provides a detailed description of our proposed holistic MDS summarization framework. We first explain how we formulate the MDS problem holistically in Section 3.1. The overall architecture of our holistic framework is shown in Figure 2, which includes holistic inference methods for summary sentence extraction in Section 3.2, and a new holistic measurement, the Subset Representative Index (SRI) in Section 3.3.

3.1 Problem Formulation

Multi-document summarization typically takes a collection of n documents $\mathcal{D} = \{D^{(1)}, \ldots, D^{(n)}\}$ as inputs. Each document contains a varying number of sentences $D^{(i)} = \{s_0^{(i)}, \ldots, s_{l_i}^{(i)}\}$, where l_i is the number of sentences in the i-th document. Let $\mathcal{S}$ be the collection of all sentences, i.e. $\mathcal{S} = D^{(1)} \cup \cdots \cup D^{(n)}$. Additionally, let $e_{i,j}$ denote the similarity score between sentence s_i and sentence s_j. Our goal is to select a representative subset of sentences $\mathcal{S}' \subset \mathcal{S}$ that maximizes the total importance of the subset while minimizing the redundancy within sentences in the subset at the same time.

3.2 Holistic Inference

Most existing approaches for unsupervised extractive summarization formulate it as an individual sentence ranking problem. They first calculate a measurement $\mathcal{M}(s_i)$ (e.g. sentence importance) for each sentence $s_i \in \mathcal{S}$ and rank all sentences in $\mathcal{S}$ accordingly. For summary inference, they directly use an *individual greedy* method that adds one sentence with the highest ranking at a time until the desired total number of summary sentences is reached. In contrast, a holistic summarization method should evaluate a subset of sentences $\mathcal{M}(\mathcal{S}')$ as a whole, then select the best subset $\mathcal{S}'$. The setting formulates the holistic summary inference into a best subset selection problem, which has exponential time complexity.

To address the exponential time complexity issue, we propose several holistic inference methods for summary sentence extraction. These methods optimize subsets of sentences using subset measurements, as opposed to the individual greedy inference method. We describe the different variants of the proposed method as below.

Holistic Greedy Method. The most straightforward way to address the exponential time complexity issue is to adopt a greedy approach. Similar to the individual greedy method, the holistic greedy method also adds one sentence at a time. However, it picks the sentence using a subset measurement that takes into account the previously selected sentences. Formally, at each step, the method selects the sentence that maximizes the following objective:

$$\text{argmax}_{s_i \in \mathcal{S} \setminus \mathcal{S}'} \mathcal{M}(\mathcal{S}' \cup \{s_i\}), \quad (1)$$

where $\mathcal{S}'$ represents the previously selected sentences.

Holistic Exhaustive Search. It is a brute-force method that considers every possible subset with the desired number of sentences. However, due to the exponential computation time, it is necessary to first filter out low-importance candidates using $\mathcal{M}(\{s_i\})$ to reduce the search space.

Holistic Beam Inference . We also propose Holistic Beam Inference which balances the trade-off between search space size and efficiency. It is a more advanced holistic inference method that adapts the beam-search decoding algorithm. We illustrate the algorithm in Algorithm 1. At each step, it considers the top-k candidate subsets, which enlarges the search space and therefore has a higher chance of finding a better subset solution compared to the holistic greedy method. Meanwhile, the algorithm has linear time complexity, making it more efficient than the holistic exhaustive search method.

3.3 Subset Representative Index

To complement the holistic inference methods, we propose a new subset measurement, Subset Representative Index (SRI), denoted as $\mathcal{M}(\mathcal{S}')$. It balances the importance measurement $\mathcal{I}(\mathcal{S}')$ and redundancy measurement $\mathcal{R}(\mathcal{S}')$.

An ideal extractive summary should select the most representative subset from a collection of the input sentences, maximizing the total non-redundant salient information passed to the user. SRI is a holistic subset measurement that balances the importance and redundancy of a subset of sentences from the source documents. Formally, we define SRI as below:

$$\mathcal{M}(\mathcal{S}') = \mathcal{I}(\mathcal{S}') - \lambda \cdot \mathcal{R}(\mathcal{S}'), \tag{2}$$

where $\mathcal{I}(\mathcal{S}')$ measures the informativeness of a set of sentences, and $\mathcal{R}(\mathcal{S}')$ measures the redundancy within the set. The parameter λ is used to control the weight of the redundancy in the overall SRI score. We detail the methods for measuring the set importance and redundancy in an unsupervised manner as follows.

Graph-Based Importance Measurement. To measure the importance of sentences, we use a graph-based approach. We construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where node $v_i \in \mathcal{V}$ represents sentence $s_i \in \mathcal{S}$, and edge $e_{i,j} \in \mathcal{E}$ represents the similarity between sentence s_i and s_j. Our proposed approach for sentence similarity score employs a combination of two methods: TF-IDF and Sentence-BERT (Reimers and Gurevych, 2019). TF-IDF is used to encode sentences with surface-form similarity, while Sentence-BERT is used to encode sentences with semantic similarity:

$$e_{i,j} = \alpha \cdot \mathbf{c}_i^\top \mathbf{c}_j + (1-\alpha) \cdot \mathbf{r}_i^\top \mathbf{r}_j, \tag{3}$$

where $\mathbf{c}_i$, $\mathbf{c}_j$, $\mathbf{r}_i$ and $\mathbf{r}_j$ are the corresponding TF-IDF features and sentence embeddings for the i-th and j-th sentences, respectively. The weight term $\alpha \in [0, 1]$ is a configurable hyperparameter to balance between statistical similarity and contextualized similarity.

Inspired from (Mihalcea and Tarau, 2004; Erkan and Radev, 2004), we define the importance of a sentence as its node centrality in the graph, which is calculated as the sum of the weights of edges connected to the node representing this sentence:

$$\mathcal{I}(s_i) = \sum_{s_j \in \mathcal{S} \setminus s_i} e_{i,j}. \tag{4}$$

Similarly, the importance of a subset of sentences is defined as the total weights between the subgraph and the remaining graph:

$$\begin{aligned}\mathcal{I}(\mathcal{S}') &= \frac{1}{|\mathcal{S}| - |\mathcal{S}'|} \sum_{s_i \in \mathcal{S}', s_j \in \mathcal{S} \setminus \mathcal{S}'} e_{i,j} \\ &\approx \frac{1}{|\mathcal{S}|} \sum_{s_i \in \mathcal{S}', s_j \in \mathcal{S} \setminus \mathcal{S}'} e_{i,j}.\end{aligned} \tag{5}$$

Since $|\mathcal{S}'|$ is usually far smaller than $|\mathcal{S}|$ in summarization tasks, we can approximate the denominator by using $|\mathcal{S}|$ directly. This way, the subset importance only takes into account the relationship of the

Algorithm 1 Holistic Beam Inference

Input: set of sentences $\mathcal{S}$, Measurement $\mathcal{M}(\cdot)$
Parameter: # summary sentences $N < |\mathcal{S}|$, beam size k
Output: the selected subset $\mathcal{S}'$

The candidate set $\mathcal{C} \leftarrow \{\varnothing\}$
for N times **do**
  The beam set $\mathcal{C}' \leftarrow \{\varnothing\}$
  for $\mathcal{X} \in \mathcal{C}$ **do**
    $\mathcal{X}' \leftarrow$ arg-top-k-max$_{s \in \mathcal{S} \setminus \mathcal{X}} \mathcal{M}(\mathcal{X} \cup \{s\})$
    for $x \in \mathcal{X}'$ **do**
      Add $\mathcal{X} \cup \{x\}$ to $\mathcal{C}'$
    end for
  end for
  $\mathcal{C} \leftarrow$ arg-top-k-max$_{\mathcal{X} \in \mathcal{C}'} \mathcal{M}(\mathcal{X})$
end for
return $\operatorname{argmax}_{\mathcal{X} \in \mathcal{C}} \mathcal{M}(\mathcal{X})$

subset with the remaining sentences, rather than considering dependencies within the subset.

Adaptive Importance Measurement. In spite of the data sparsity issue in MDS, the Single Document Summarization (SDS) task has abundant high-quality labeled data (Hermann et al., 2015; Narayan et al., 2018; Cohan et al., 2018). We propose a method called adaptive importance measurement, which adapts SDS data for MDS importance measurement. This method utilizes the labeled data from SDS to train a model for predicting the importance of sentences in MDS.

In the adaptive setting, we fine-tune the BERT (Devlin et al., 2018) to a sentence importance scorer on SDS datasets, and then adapt the fine-tuned model to the target MDS datasets. Specifically, we first calculate the normalized salience of a sentence as:

$$f(s_i) = \mathbf{v}^\top \tanh(\mathbf{W}_1 \mathbf{r}_i),$$
$$\text{salience}(s_i) = \frac{f(s_i)}{\sum_{s_j \in D} f(s_j)}, \quad (6)$$

where $\mathbf{W}$ is a trainable weight, and $\mathbf{r}_i$ is the contextualized representation of sentence s_i. Then, we fine-tune BERT to minimize the following loss:

$$R(s_i) = \text{softmax}(\text{ROUGE}(s_i)),$$
$$\mathcal{L} = -\sum_{D} \sum_{s_i \in D} R(s_i) \log \text{salience}(s_i). \quad (7)$$

The fine-tuned BERT can be directly adapted to the MDS datasets and calculate the adaptive importance measurement for sentences.

Redundancy Measurement. The redundancy measurement for a subset of sentences $\mathcal{S}'$ is defined as the total similarity score of each sentence with its most similar counterpart. This measurement captures the degree of overlap between the sentences in the subset, indicating the level of redundancy present in the selected sentences:

$$\mathcal{R}(\mathcal{S}') = \sum_{s_i \in \mathcal{S}'} \max_{s_j \in \mathcal{S}' \setminus \{s_i\}} e_{i,j}. \quad (8)$$

Overall, we can calculate SRI in both unsupervised and adaptive manners. Our holistic framework extracts summaries as a whole with the holistic inference method, which is guided by SRI to measure the importance and redundancy of a subset of sentences. This approach allows us to balance the importance and redundancy of a summary, making it more informative and coherent.

4 Experiments

In this section, we provide details on our experimental setup, including the datasets, evaluation metrics, baselines, and implementation details (Section 4.1). We then present the results of our model on benchmark MDS datasets in both unsupervised (Section 4.2) and adaptive (Section 4.3) settings.

4.1 Experimental Setting

Dataset. We evaluate our unsupervised method on benchmark multi-document summarization datasets. Particularly, we use MultiNews (Fabbri et al., 2019), WikiSum (Liu et al., 2018), DUC-04 (Over and Yen, 2004), and TAC-11 (Dang and Owczarzak, 2008) datasets. MultiNews is collected from a diverse set of news articles on newser.com. It is a large-scale dataset containing reference summaries written by professional editors. WikiSum is another large-scale dataset that provides documents and summaries from Wikipedia webpages where the documents come from the reference webpages of Wikipedia articles and top-10 Google searches, and the summaries are the lead section of the Wikipedia articles. We use the top-40 high-ranked paragraphs for the document inputs following (Liu and Lapata, 2019a).

For summary extraction, we use the average number of reference sentences: 10 and 5, respectively on MultiNews and WikiSum. For the DUC and TAC datasets, the task is to generate a succinct summary of up to 100 words from a set of 10 news articles. We report results on DUC-04 and TAC-11, which are standard test sets used in previous studies (Hong et al., 2014; Cho et al., 2019a). DUC-03 and TAC-08/09/10 are used for the validation set to tune hyper-parameters. For adaptive setting, we fine-tune BERT on single document summarization dataset CNN/DailyMail (Hermann et al., 2015) and directly adapt to MDS test sets. Table 1 shows the statistics of the datasets in detail.

Evaluation Metrics. The extracted summaries are evaluated against human reference summaries using ROUGE (Lin, 2004)[1] for the summarization quality. We report ROUGE-1, ROUGE-2, ROUGE-SU4, and ROUGE-L[2] that respectively measure the overlap of unigrams, bigrams, skip bigrams with a maximum distance of 4 words, and the longest

[1] w/ options -n 2 -m -w 1.2 -c 95 -r 1000 -l 100 for DUC/TAC

[2] Due to some legacy issues, some baselines report the original ROUGE-L, others report ROUGE-Lsum.

Dataset	# test	# ref.	avg.w/doc	avg.w/sum
DUC-04	50	4	4,636	109.6
TAC-11	44	4	4,696	99.7
Multi-News	5,622	1	2,104	264.7
Wikisum	38,205	1	2,800	139.4
CNNDM(SDS)	11,489	1	766.1	58.2

Table 1: Detailed statistics of four multi-document datasets. #test denotes the number of document clusters in the test set, #ref denotes the number of reference summaries, avg.word(doc) denotes the average number of words in the source document cluster, avg.word(sum) denotes the average number of words in the ground truth summary.

common sequence between extracted summary and reference summary. To align with previous works, we report R-1, R-2, R-L for Multinews and Wikisum datasets, and R-1, R-2, R-SU4 for DUC and TAC datasets. For all baseline methods, we report ROUGE results from their original papers if available or use results reported in (Cho et al., 2019a; Liu et al., 2021). We also report the measure of diversity for the generated summaries by calculating a unique n-gram ratio (Xiao and Carenini, 2020; Peyrard et al., 2017) defined as:

$$\text{uniq } n\text{-gram ratio} = \frac{\#\text{ uniq-}n\text{-gram}}{\#n\text{-gram}} \quad (9)$$

Baselines. We compare our methods with strong unsupervised summarization baselines. In particular, *MMR* (Carbonell and Goldstein, 1998) combines query relevance with information novelty in the context of summarization. *LexRank* (Erkan and Radev, 2004) computes sentence importance based on eigenvector centrality in a graph representation of sentences. *TextRank* (Mihalcea and Tarau, 2004) adopts PageRank (Page et al., 1999) to compute node centrality recursively based on a Markov chain model. *SumBasic* (Vanderwende et al., 2007) is an extractive approach assuming words frequently occurring in a document cluster are more likely to be included in the summary. *KLSum* (Haghighi and Vanderwende, 2009) uses a greedy approach to add a sentence to the summary to minimize the KL divergence. *PRIMERA* (Xiao et al., 2022) is a pyramid-based pre-trained model for MDS that achieves state-of-the-art performance. We compare it under its zero-shot setting.

Implementation Details. We run all experiments with 88 Intel(R) Xeon(R) CPUs. We combine the surface indicator based on TF-IDF and contextualized embeddings. We treat each document cluster as a corpus and each sentence as a document when calculating the TF-IDF scores. We employ the pre-trained sentence-transformer (Reimers and Gurevych, 2019) and extract sentence representations using a checkpoint of 'all-mpnet-base-v2'.

The graph edges with low similarity are treated as disconnected to emphasize the connectivity of the graph and avoid noisy edge connections. We keep a threshold $\tilde{e}$ for edge weights such that edges with similarity scores smaller than $\tilde{e}$ will be set to 0. Here $\tilde{e}$ is controlled by a hyper-parameter to be tuned according to datasets. The final representation of edge weight between two sentences (s_i, s_j) is

$$e_{i,j} = \max(\text{sim}(s_i, s_j) - \tilde{e}, 0), \quad (10)$$

where $\tilde{e} = \min(e) + \theta\ (\max(e) - \min(e))$ is the threshold controlled by hyper-parameter θ. For exhaustive search, we filter out the sentences with low centrality and only keep the top 15 sentences at inference.

All hyper-parameters are tuned on validation sets on MultiNews and WikiSum and training sets on DUC and TAC. The best parameters are selected based on the highest R-1 score. More specific, for the balancing factor λ in SRI, we use $\{2^{-13}, 2^{-7}, 2^{-4}, 2^{-6}\}$ on DUC, TAC, MultiNews and WikiSum dataset. For α that weighted the contributions of TF-IDF and contextualized sentence similarity, we use 0.9 on News domain datasets and 0.8 on the WikiSum dataset. The edge weight threshold θ is $\{0, 0, 0.1, 0.1\}$ for DUC, TAC, MultiNews and WikiSum. As for beam search, we use beam size $\{4, 4, 4, 3\}$ on the corresponding datasets.

4.2 Unsupervised Summarization Results

The unsupervised summarization results on four benchmark MDS datasets are shown in Table 2.

The summarization performance of our method outperforms strong unsupervised baselines. Note that MultiNews and WikiSum datasets provide abundant training samples and contain shorter input than the DUC or TAC datasets. Our method performs better than the pre-trained model, *PRIMERA* with a zero-shot setting. Compared to the baseline (Sent. greedy) that extracts sentences solely based on importance, balancing diversity with SRI boosts performance by a large margin.

For the DUC-04 and TAC-11 datasets, our proposed methods outperform unsupervised baselines by a large margin. It demonstrates that balancing the summary informativeness and diversity during

System	*DUC-04*			*TAC-11*			*MultiNews*			*WikiSum*		
	R-1	**R-2**	**R-SU**	**R-1**	**R-2**	**R-SU**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L***
Unsupervised Systems												
LEAD	30.77	8.27	7.35	32.88	7.84	11.46	39.41	11.77	14.51	37.63	14.75	33.76
MMR (1998)	30.14	4.55	8.16	31.43	6.14	11.16	38.77	11.98	12.91	31.22	10.24	22.48
LexRank (2004)	34.44	7.11	11.19	33.10	7.50	11.13	38.27	12.70	13.20	36.12	11.67	22.52
TextRank (2004)	33.16	6.13	10.16	33.24	7.62	11.27	38.44	13.10	13.50	23.66	7.79	21.23
SumBasic (2007)	29.48	4.25	8.64	31.58	6.06	10.06	-	-	-	-	-	-
KLSumm (2009)	31.04	6.03	10.23	31.23	7.07	10.56	-	-	-	-	-	-
PRIMERA (2022)	35.10	7.20	17.90	-	-	-	42.00	13.60	20.80	28.00	8.00	18.00
Individual. Greedy	34.81	7.85	11.37	34.42	8.10	11.25	40.48	13.49	16.14	37.24	10.29	32.77
Our Methods												
SRI+beam	**36.84**	**8.37**	12.28	**35.37**	**8.49**	**11.73**	**44.22**	**14.63**	18.61	38.94	15.23	34.12
SRI+exh	36.70	**8.37**	**12.31**	35.19	8.31	11.34	43.16	14.58	18.00	**39.26**	**16.15**	**34.19**

Table 2: ROUGE-F1 scores on four datasets under the unsupervised setting. Best unsupervised results are bold. For a fair comparison, we report R-L on Multinews and R-Lsum (See et al., 2017) for WikiSum and limit summaries to 100 words on DUC-04 and TAC-11. R-L are marked with * if reporting ROUGE- Lsum numbers.

Method	**Fluent**	**Informative**	**Faithful**	**Overall**
MMR	3.2	3.5	**4.7**	3.2
PRIMERA	**4.3**	2.5	3.3	3.3
SRI	3.8	**4.3**	**4.7**	**4.0**

Table 3: Human evaluation results on a scale of 1-5.

System	*DUC-04*			*MultiNews*		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
Adaptive Systems						
BART(2019)	24.1	4.0	15.3	27.3	6.2	15.1
BART (CNNDM)	29.4	6.1	16.2	36.7	8.3	17.2
PEGASUS (2020)	32.7	7.4	17.6	32.0	10.1	16.7
PEGASUS(CNNDM)	34.2	7.5	17.4	35.1	11.9	18.2
LED(2020)	16.6	3.0	12.0	17.3	3.7	10.4
PRIMERA (2022)	35.1	7.2	17.9	42.0	13.6	20.8
Our Systems						
SRI+beam (graph)	36.8	8.4	16.4	44.2	**14.6**	18.6
SRI+beam (CNNDM)	**36.9**	**8.6**	**18.5**	**44.6**	14.3	**21.1**

Table 4: ROUGE-F1 results on DUC-04 and Multinews datasets under the adaptive setting. Models adapted from CNN/DailyMail dataset are marked in the bracket.

the sentence extraction process is crucial for better summary quality. Note that the input length of DUC/TAC datasets is extremely long spanning an average of 180 sentences. These long input easily exceeds the input capacity of transformer-based models possibly resulting in information loss from documents. The proposed methods on the other hand process documents regardless of the input length or formats (SDS or MDS). Also, our unsupervised methods have the advantage of processing datasets with small training data. The supreme performances on datasets with different input lengths and low-resource data illustrate the effectiveness of our methods. To further verify the model performance, we also conduct a human evaluation by experts on a scale of 5. The results shown in Table 3 also prove our method outputs better summaries in unsupervised setting.

4.3 Adptive Summarization Results

The experimental results under the adaptive setting are shown in Table 4. Compared to large pre-trained generation model (BART) and other task-specific pre-trained summarization models (PEGASUS, PRIMERA), our framework shows strong performance when adapting from a single document summarization dataset. We also notice fine-tuning on single document summarization corpus improves the performance of all pre-trained models, but still, our framework achieves the best results under the adaptive setting.

5 Analysis

5.1 Summary Diversity

Other than summary quality, we also test the effectiveness of our SRI in terms of the diversity of the output summaries. We present the unique n-gram ratios of output summaries under unsupervised and adaptive settings and the reference summary on the TAC-11 dataset in Figure 3. According to the results, our framework is extremely effective in reducing summary redundancy and increasing summary diversity under both unsupervised and adaptive settings.

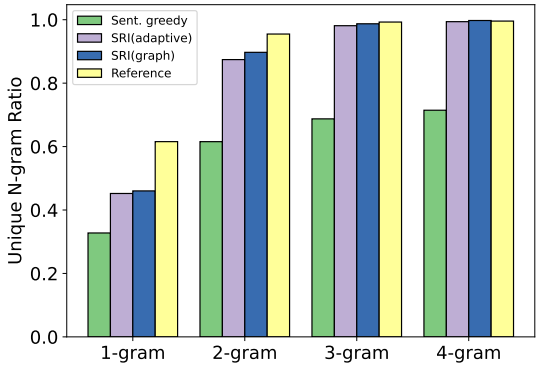

Figure 3: Unique n-gram ratios ($n = 1, 2, 3, 4$) of the output summary by different methods on TAC-11.

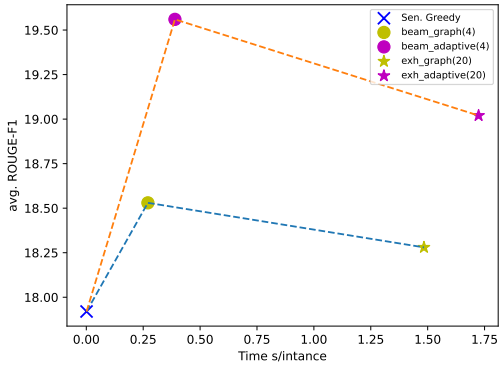

Figure 5: Efficiency vs. average ROUGE (w/o word limit) scores of different inference methods on TAC-11.

Compared to the ROUGE-F1 results, holistic inference with importance-diversity balancing measurement SRI increases both summary quality and diversity at the same time. The results suggest that considering summary diversity is beneficial in extractive summarization, especially in redundant cases like MDS and long document summarization. Our finding also verifies the crucial rule of effective modeling of sentence importance and similarity.

5.2 Hyperparameter Sensitivity

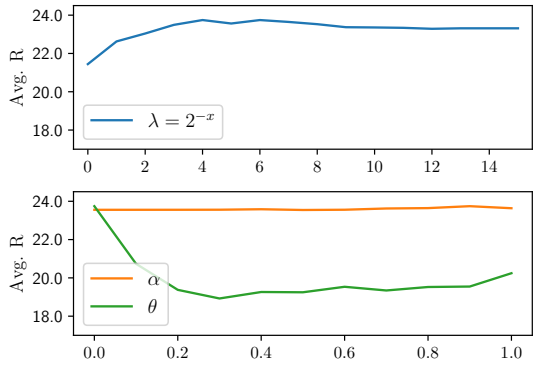

Figure 4: Average ROUGE-F1 (w/o word limit) results with different hyperparameter values on TAC-11.

To test the robustness of our proposed approaches, we study the hyperparameter sensitivity of our proposed methods. The results are shown in Figure 4. The first plot shows the impact of balancing factor λ in SRI. The second plot shows the impact of α, which balances the contextualized and TF-IDF sentence embedding and the edge weight threshold. The results show that our methods are relatively stable towards the hyperparameter values and could be easily adapted to unseen datasets.

Beam Size	**2**	**3**	**4**	**5**	**6**	**7**	**8**
ROUGE-1	33.43	33.65	33.62	**33.76**	34.72	33.64	33.67
ROUGE-2	7.71	**8.00**	7.87	7.93	7.84	7.86	7.85
ROUGE-L*	28.74	29.03	28.99	**29.10**	29.01	28.94	29.01

Table 5: ROUGE-F1 (w/o word limit) results of SRI-beam with different beam sizes on TAC with $\lambda = 0.125$.

5.3 Inference Approaches Analysis

We also compare the efficiency and effectiveness of different inference methods. As in Figure. 5, we compare sentence-level greedy search, set-level greedy search, set-level beam search (beam size = 4), and set-level exhaustive search with pre-filtering as inference methods for both unsupervised and adaptive settings. We pick the filter size of 20 here since the search space without filtering $C(N, K)$ is extremely large. According to the results, all set-level inference methods outperform the sentence-level methods. This suggests that extracting summaries at a set level (holistic) is optimal over the common sentence-level setting that extracts sentences individually. The finding is also consistent with the inherent performance gap between sentence-level and holistic extractors in (Zhong et al., 2020).

Moreover, we realize the set-level beam search and set-level exhaustive search achieve the comparable best performance. However, set-level beam search speed-wise is much more efficient than set-level exhaustive search. We also show the effect of different beam sizes in Table 5. The results indicate that a reasonably small beam size achieves the best ROUGE results, which are both effective and efficient. To conclude, set-level beam search with SRI shows the best overall performance.

6 Conclusion

This paper proposes a holistic framework for unsupervised multi-document extractive summarization. Our framework incorporates the holistic beam search inference methods and SRI, a holistically balanced measurement between importance and diversity. We conduct extensive experiments on both small and large-scale MDS datasets under both unsupervised and adaptive settings and the proposed method outperforms strong baselines by a large margin. We also find that balancing summary set importance and diversity benefits both the quality and diversity of output summaries for MDS.

Limitations

The proposed framework in this paper is mainly designed for low-resource scenarios without gold summaries for multi-document summarization. Adapting the framework for a supervised setting requires further investigation. Recently, large language models (LLM) like ChatGPT have shown strong zero-shot summarization ability, which may raise doubt about the necessity of unsupervised summarization methods.

However, LLM suffers from the hallucination problem and MDS may exceed its input limit (e.g. 4,696 words for TAC) than the input limit of ChatGPT (500-word/4,000-character). In contrast, unsupervised summarization methods can tackle input of arbitrary length and have a faster inference speed than ChatGPT when processing long input documents. In addition, a recent study (Zhang et al., 2023c) shows that ChatGPT's extractive summarization performance is still inferior to existing supervised systems in terms of ROUGE scores.

Ethical Consideration

Our proposed framework forms summary by directly extracting sentences from source documents. Therefore, the extracted summary may be incoherent or contain unfactual co-references. In addition, the extracted summary will keep biased contents from the source sentences, if any.

References

Chenxin An, Ming Zhong, Zhiyong Wu, Qin Zhu, Xuanjing Huang, and Xipeng Qiu. 2022. Colo: A contrastive learning based re-ranking framework for one-stage summarization. *arXiv preprint arXiv:2209.14569*.

Diego Antognini and Boi Faltings. 2020. Gamewikisum: a novel large multi-document summarization dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6645–6650, Marseille, France. European Language Resources Association.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

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

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019a. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1027–1038, Florence, Italy. Association for Computational Linguistics.

Sangwoo Cho, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2019b. Multi-document summarization with determinantal point processes and contextualized representations. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 98–103, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 update summarization task. In *Proceedings of Text Analysis Conference (TAC)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6244–6254.

Alex Kulesza and Ben Taskar. 2012. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. *arXiv preprint arXiv:1808.06218*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Improving unsupervised extractive summarization with facet-aware modeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1685–1697.

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

Hui Lin, Jeff Bilmes, and Shasha Xie. 2009. Graph-based submodular selection for extractive summarization. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 381–386. IEEE.

Jingzhou Liu, Dominic JD Hughes, and Yiming Yang. 2021. Unsupervised extractive text summarization with distance-augmented sentence graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2313–2317.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam M. Shazeer. 2018. Generating wikipedia by summarizing long sequences. *ArXiv*, abs/1801.10198.

Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

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

Shulei Ma, Zhi-Hong Deng, and Yunlun Yang. 2016. An unsupervised multi-document summarization framework based on neural document model. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1514–1523.

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

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*.

Paul Over and James Yen. 2004. An introduction to DUC-2004. *National Institute of Standards and Technology*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

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

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

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

Jun Suzuki and Masaaki Nagata. 2017. Cutting-off redundant repeating generations for neural abstractive summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 291–297, Valencia, Spain. Association for Computational Linguistics.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6209–6219, Online. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Wen Xiao and Giuseppe Carenini. 2020. Systematically exploring redundancy reduction in summarizing long documents. *arXiv preprint arXiv:2012.00052*.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive text summarization. *arXiv preprint arXiv:1910.14142*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2022. HEGEL: Hypergraph transformer for long document summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10167–10176, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Contrastive hierarchical discourse graph for scientific document summarization. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 37–47, Toronto, Canada. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. Diffusum: Generation enhanced extractive summarization with diffusion. *arXiv preprint arXiv:2305.01735*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023c. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. *arXiv preprint arXiv:1906.03508*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.