

Disentangling Specificity for Abstractive Multi-document Summarization

Congbo Ma¹, Wei Emma Zhang², Hu Wang², Haojie Zhuang², Mingyu Guo²

¹Macquarie University, Australia ²The University of Adelaide, Australia

congbo.ma@mq.edu.au, {wei.e.zhang, hu.wang, haojie.zhuang, mingyu.guo}@adelaide.edu.au

Abstract—Multi-document summarization (MDS) generates a summary from a document set. Each document in a set describes topic-relevant concepts, while per document also has its unique contents. However, the document specificity receives little attention from existing MDS approaches. Neglecting specific information for each document limits the comprehensiveness of the generated summaries. To solve this problem, in this paper, we propose to disentangle the specific content from documents in one document set. The document-specific representations, which are encouraged to be distant from each other via a proposed orthogonal constraint, are learned by the specific representation learner. We provide extensive analysis and have interesting findings that specific information and document set representations contribute distinctive strengths and their combination yields a more comprehensive solution for the MDS. Also, we find that the common (i.e. shared) information could not contribute much to the overall performance under the MDS settings. Implementation codes are available at <https://github.com/congboma/DisentangleSum>.

Index Terms—Multi-document summarization, Deep neural network, Transformer

I. INTRODUCTION

Multi-document summarization (MDS) is an important task in the natural language processing [1]. Processing the source documents by flat-concatenating them into a mega document is one way to solve MDS tasks [2]–[4]. But the purpose of MDS is to treat each document as a standalone unit, digging out the connections and differences between documents and generating informative and comprehensive summaries. Targeting these issues, some researchers tried to establish the connections not only at word-level relations but also sentence, paragraph and document levels. They employ hierarchical Transformer structures [5]–[8] to forge connections among documents. The high-level Transformer encodes the paragraph representations from different documents. Besides, some existing works incorporated graph information [6], [9], [10] to build connections among documents. However, these methods are not specifically designed for extracting specific features and therefore they ignore the specific information contained in each document in a document set.

Nonetheless, the extraction of specific information is crucial with the following reasons: (1) In a collection of documents, each document contains not only the common information but also has specific contents that distinguish it from other documents. These specific information contain unique facts, viewpoints, and details [2]. Extracting these specific details enhance the comprehensiveness of the resulting summary. Additionally, some essential information may be exclusive to

a particular document, yet it plays a pivotal role in obtaining a comprehensive grasp of the entire document set. Therefore, a high-quality MDS summary should not only be able to capture document commonality but can also comprehensively consider the specific information from each document, covering various dimensions to meet the user’s demand for a comprehensive understanding of the documents [11]. (2) Focusing on the extraction of specific information helps reduce redundancy, rendering the summary more concise and informative. Clustering-based MDS methods [9], [12], [13] can be used to group similar sentences or pieces of information and remove redundancy. After removing the redundant information, the remaining information in each document can be viewed as implicitly specific information. However, the specificity within these remaining information cannot be explicitly guaranteed to be distinctive between documents.

In order to address this issue, our intuition is not only to capture the overall information in a document set but also to distinguish the specificity of each document and learn representations of document specificity which will be considered in the summary generation process. To this end, we propose DisentangleSum — a simple yet effective summarization model that disentangles document uniqueness with a set of document-specific representation learners. In order to optimize the learning of specific representations, we further propose an orthogonal constraint to encourage the specific representations obtained from a pair of documents to be distinctive from each other. Based on the constraint, we design an objective function that can transform the quadratic increment of the losses between each of the paired documents into linear to cope with a large number of documents in a set. We summarize our contributions as follows:

- We present DisentangleSum, an innovative MDS model that is capable of disentangling specific information from each document in a set, leading to more comprehensive summary generation. To the best of our knowledge, we are the first to consider the specific information for deep learning based MDS task.
- To incentivize the document-specific learner to retain document specificity information, we propose an orthogonal constraint. This constraint encourages the document-specific representation vectors to align vertically with each other, ensuring a semantic separation between them.
- Experimental results on two MDS datasets demonstrate

the effectiveness of DisentangleSum. We additionally offer comprehensive analyses from multiple perspectives to investigate the underlying mechanisms of DisentangleSum and circumstances of the proposed model can work.

II. RELATED WORKS

Mining document set representations. One way to process MDS input is to concatenate documents into a single mega document, utilizing overall document set representations [2]–[4]. This concatenation neglected the relative importance of documents, reordering the documents in a document set according to their importance makes it easier for the summarization model to learn [4]. However, Methods deal with mega documents do not explicitly consider interconnections and distinctions among individual input documents [5].

Mining document relation representations. Researchers have focused on mining relationships among documents in the same set using different techniques. These include extracting graph representations [6], [10], [14], leveraging document-level positional relations [15], and performing entity extraction [16]. These methods incorporate domain-specific knowledge or semantic relationships among the source documents for better performance. Additional approaches to document relation mining include hierarchical Transformer architectures [5] and attention mechanisms with different granularity representations [7]. However, the existing methods mentioned above overlook the distinctiveness of individual documents within a document set, inevitably leading to incomprehensive summarization. To fill this research gap, we propose the DisentangleSum model, which extracts document-specific representations for comprehensive summary generation.

III. METHODOLOGY

In this section, we provide an overview of the proposed model (Figure 1), DisentangleSum, by describing how to incorporate document disentangling specificity representation learning into the summarization framework. We propose the orthogonal constraint applied during the training of document-specific representations.

A. Problem Formulation

In MDS, each document set can have a varying number of documents. For illustration purposes, let's consider a document set $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \dots, \mathbf{d}_N)$ consisting of N input documents related to a specific topic or sharing common information. In our approach, we utilize the specific encoder $\psi_i(\cdot; \Theta)$ for document \mathbf{d}_i , where Θ represents the learnable parameters. These specific encoders generate specific representations \mathbf{S}_i for each document, and collectively, they form the specific representations \mathbf{S} for the entire document set. Additionally, we employ a document-set encoder $\phi(\cdot; \Omega)$ with learnable parameters Ω to obtain document-set representations \mathbf{F} . The target is to generate a concise summary output \mathbf{O} that synthesizes all important contents from input documents by considering both specific representations \mathbf{S} and document-set representations \mathbf{F} .

B. Document Specific Representation Learner

In a document set, the specific representation learner aims to identify the specific information within each document. To achieve this, we introduce a specific encoder to encode document \mathbf{d}_i in the same document set \mathbf{D} :

$$\mathbf{S}_i = \psi_i(\mathbf{d}_i; \Theta), \quad (1)$$

Under the setting of MDS, the number of input documents can vary within a document set. To address this variability, we propose a design where the learnable parameters, denoted as Θ , are shared across a set of N specific encoders instead of assigning a separate specific encoder to each document in the set. The rationale behind this approach stems from the fact that documents with identical indexes in different document sets are unrelated in content. Consequently, maintaining multiple separate specific encoders for each indexed document is not reasonable. Subsequently, we concatenate these specific representations to obtain the overall specific representations of a document set:

$$\mathbf{S} = \mathbf{S}_1 \oplus \mathbf{S}_2 \oplus \mathbf{S}_3 \oplus \dots \oplus \mathbf{S}_N, \quad (2)$$

\oplus is the concatenation operation in this paper. To enable sufficient expressive power for representations to be decoded, we also obtain the document-set representations \mathbf{F} out of a document set \mathbf{D} by:

$$\mathbf{F} = \phi(\mathbf{D}; \Omega), \quad (3)$$

Next, we combine the document-set representations and specific representations by performing an element-wise addition and decoding them into summarization outputs:

$$\mathbf{O} = \zeta(\alpha \cdot \mathbf{F} + \mathbf{S}; \Lambda), \quad (4)$$

Here, α serves as a trade-off factor to control the weight balancing between the document-set representations and specific representations. The dimension of \mathbf{F} and \mathbf{S} are both equal to the length of a document set. The decoder function $\zeta(\cdot; \Lambda)$, parameterized by Λ , is responsible for decoding the intermediate representations into concise summaries.

C. Orthogonal Constraint within the Training of Document Specific Representations

To guide specific representations learning, we impose an orthogonal constraint between pairs of specific representations \mathbf{S}_i and \mathbf{S}_j . The document-specific loss, which promotes dissimilarity between specific representations, is defined as:

$$L_{spec} = \sum_i \sum_j \left\| \mathbf{S}_i^\top \mathbf{S}_j \right\|_F^2, \quad (5)$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm. To encourage dissimilarity between specific representations, we aim for a smaller inner product between each pair of specific representation vectors, promoting orthogonality. This ensures distinctiveness among specific representations within the same document set. As the specific encoder learns, it captures each document's unique essence, thereby retaining specific content. However,

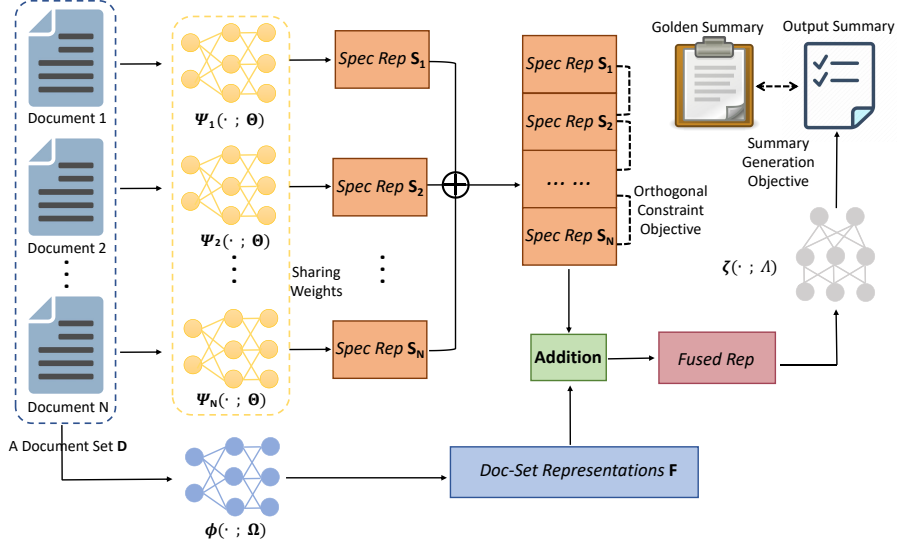


Fig. 1: The overall framework of the proposed DisentangleSum model.

when a document set contains more than two documents, the computation of specific representation objectives between every documents pair grows quadratically. To address this, we introduce a circle-paired loss objective function, reducing complexity from quadratic to linear. Formally, we have:

$$L_{spec} = \sum_{i=1}^N L_{spec}^i, \quad (6)$$

$$L_{spec}^i = \begin{cases} \|\mathbf{S}_i^\top \mathbf{S}_{i+1}\|_F^2 & i \neq N \\ \|\mathbf{S}_N^\top \mathbf{S}_1\|_F^2 & i = N \end{cases}, \quad (7)$$

The objective function calculates the specific representation costs between each document and the subsequent document in the set, with the last document computed against the first one.

D. Overall Objectives

The proposed framework aims to train a high-quality summarization model that incorporates specific representations from each document. This is achieved through two key components: an orthogonal constraint for distinct document representations and a supervised cross-entropy loss concerning golden summaries:

$$L_{total} = L_{gen} + \beta \cdot L_{spec}, \quad (8)$$

$$L_{gen} = - \sum_{k=1}^M \eta(\hat{\mathbf{O}}_k, \mathbf{O}_k) \log(p(\mathbf{O}_k)), \quad (9)$$

where β is a balance factor, $p(\mathbf{O}_k)$ is one shard¹ of the predictive summary from the DisentangleSum model. $\hat{\mathbf{O}}_k$ denotes corresponding true labels. M represents the number of shard within the generated summary. The calculation of $\eta(\cdot, \cdot)$

in a summarization task is different from other tasks such as text classification. $\eta(\cdot, \cdot)$ indicates the evaluation function between prediction and ground truth, the widely used ROUGE evaluation are adopted here.

IV. EXPERIMENTS

A. Datasets & Evaluation Metrics & Baselines

We assess the effectiveness of the proposed method on Multi-News [2] and Multi-XScience [17] datasets which satisfy that documents contain the specific information in a document set. Both datasets are truncated to 500 tokens. We use standard summarization evaluation metrics ROUGE² [18] BERTScore³ [19] Additionally, we employ the coverage score [20] to quantify the amount of information retained in the generated summaries compared to input documents. We compare with the following strong baselines: *LexRank* [21], *TextRank* [22], *MMR* [23], *BRNN*, *Vanilla Transformer (VT)* [24] and its variant *CopyTransformer (CT)*, *Pointer-Generator (PG)* [25], *Hi-MAP* [2], *Hierarchical Transformer (HierTrans)* [5], *SummPip* [14], *SAGCopy* [26], *HeterGraphSum (HGS)* [10], *Highlight-Transformer (HiTrans)* [27], *DocLing* [15].

B. Implementation Details

During model training, the initial learning rate is set to 2. The training strategy involves a warm-up phase for the first 8,000 steps, followed by multi-step learning rate reduction. The batch size is set to 4,096, and the models are trained for 20,000 steps using the *Adam* optimizer. Both the encoder and decoder consist of four transformer layers, and positional encoding is applied. The dropout rate is 0.2. The trade-off factor for specific representations (α) is set to 0.01, and the trade-off factor for specific loss (β) is set to 0.001. The word

¹Following the implementation in <https://github.com/Alex-Fabbri/Multi-News/blob/master/code/OpenNMT-py-baselines/onmt/loss.py>, shards are segments when computing losses.

²The parameters of ROUGE are -c 95 -2 -1 -U -r 1000 -n 4 -w 1.2 -a -m.

³The model type of BERTScore is bert-base-uncased.

Models	Multi-News	Multi-XScience
VT	19.76	18.98
CT	46.78	15.82
DocLing	49.62	20.10
DisentangleSum	52.99	22.68

TABLE I: Coverage score comparison.

Models	R-1	R-2	R-SU	BS
LexRank	37.92	13.1	12.51	0.83
TextRank	39.02	14.54	13.08	0.83
MMR	42.12	13.19	15.63	0.84
BRNN	38.36	13.55	14.65	0.83
VT	25.82	5.84	6.91	0.80
CT	42.98	14.48	16.91	0.84
PG	34.13	11.01	11.58	0.83
Hi-MAP	42.98	14.85	16.93	0.83
HierTrans	36.09	12.64	12.55	0.84
SummPip	42.29	13.29	16.16	0.84
SAGCopy	43.98	15.21	17.65	-
HGS	43.62	14.99	17.29	0.85
HiTrans	44.62	15.57	18.06	-
DocLing	44.35	15.04	17.97	0.85
DisentangleSum	45.95	16.32	19.23	0.85

TABLE II: Performance comparison on the Multi-News.

embedding size for source documents is set to 512 dimensions. We conduct all the experiments on one NVIDIA 3090 GPU with one Intel i9-10900X CPU upon Ubuntu 22.04.3 LTS Operation System. For the minimum and maximum lengths of generated summaries, Multi-News has 200 and 300 words, while Multi-XScience has 110 and 300 words.

C. Main Results

This section is designated for validating the model’s effectiveness from three perspectives: (1) verifying the comprehensiveness of the generated summaries; (2) evaluating the overall performance through automated evaluations; (3) evaluating human feedback on specific information extraction, comprehensive, coherence and relevance.

1) *Coverage Score*: We conduct a comparison based on coverage score (Table I) among three Transformer-based models: Vanilla Transformer, CopyTransformer, and DocLing. Despite sharing a similar structure, these models lack consideration for document specificity. Coverage score measures the percentage of words in the generated summary that come directly from the source documents. A higher score indicates that the model can generate summaries with richer information from the source documents. From the results, disentangleSum outperforms counterparts by a large margin, achieving the highest coverage score on both datasets. This indicates its ability to generate more comprehensive summaries, preserving substantial information from the original documents.

2) *Overall Performance*: Table II and Table III⁴ shows the DisentangleSum model receives outstanding performance in

⁴The code for SAGCopy [26] and HiTrans [27] is not publicly available and they did not provide results on the Multi-XScience dataset. Additionally, since MultiX-Science does not contain labels for extractive summarization, hindering the the HeterGraphSum [10] from being implemented on it.

Models	R-1	R-2	R-SU	BS
LexRank	31.31	5.85	9.13	0.83
TextRank	31.15	5.71	9.07	0.84
MMR	30.04	4.46	8.15	0.83
BRNN	27.95	5.78	8.43	0.83
VT	28.34	4.99	8.21	0.82
CT	26.92	4.92	7.50	0.83
PG	30.30	5.02	9.04	0.84
Hi-MAP	30.41	5.85	9.13	0.81
HierTrans	25.31	4.23	6.64	0.83
SummPip	29.66	5.54	8.11	0.82
DocLing	30.93	6.06	9.57	0.84
DisentangleSum	31.81	5.90	9.88	0.84

TABLE III: Performance comparison on the Multi-XScience.

Models	Speci	Compr	Coher	Relev
VT	1.67	2.28	2.13	2.46
CT	2.33	2.70	2.27	2.78
DocLing	2.89	3.10	2.78	2.93
DisentangleSum	3.16	3.87	3.21	3.13

TABLE IV: Human evaluation results on the Multi-News dataset.

most of the cases. On the Multi-News dataset, DisentangleSum outperforms the second-best model, attaining 1.6 improvement on ROUGE-1, 1.28 improvement on ROUGE-2, and 1.26 improvement on ROUGE-SU. Particularly, the ROUGE-SU score received 7% improvement over the second-best model. Similarly results are shown on Multi-XScience data as well. The superior results can be consistently gained because the proposed DisentangleSum model has been enhanced with the capability to capture both the document set and document-specific features, leading to a better summary generation.

3) *Human Evaluation*: We conduct human evaluations to assess summary quality in terms of Specificity (Speci), Comprehensiveness (Compr), Coherence (Coher), and Relevance (Relev), aiming to detect diverse viewpoints from multiple documents. Three Ph.D students in NLP examine the performance of four models using 50 randomly sampled source documents from the Multi-News dataset, scoring from one to five (one = very bad; five = very good). The final scores (shown in Table IV), averaged across different cases and raters, consistently favor DisentangleSum across all four metrics. An example (In Table V) from the MultiNews dataset [2] discusses how to discourage smoking, with each document covering the topic from different perspectives. Doc #1 indicates the ugliest colour serves an important purpose, Doc #2 lists the statistics related to smoking, and Doc #5 discusses smoking from a legal point of view. Existing works provide summaries that miss some specific information. For example, the summary generated by the DocLing [15] model fails to include the statistical information presented, resulting in the omission of important specific details from the source documents. These indicate the summaries generated by our model can cover more specific information from the source documents, and exhibit better coherence and relevance.

Doc Set	<p>Doc #1: the world’s ugliest color has been described as “death,” “dirty” and “tar,” but this odious hue is serving an important purpose: discouraging smoking. ... the agency was hired by the australian government to find a color that was so repugnant that if it was on tobacco products, it would dissuade ...</p> <p>Doc #2: ... changes to australia ’ s duty free tobacco allowance smoking prevalence rates abs national aboriginal and torres strait islander social survey, 2014-15 the proportion of aboriginal and torres strait islander people aged 15 years and over who were daily smokers was 38.9 % in 2014-15 , down from 44.6 % in 2008 and 48.6 % in 2002 ...</p> <p>Doc #5: ... in may, previously passed legislation will go into effect requiring all packs of cigarettes to be standardized. tobaccos products will be stripped of brightly colored branding and replaced with a sludge-like color . . but does the stripped-down , “ ugly ” packaging really reduce smoking ...</p>
VT	–australia’s news agency says it’s time to get rid of certain types of (unk)products. the australian government has approved a ban on (unk) products , which include ...
CT	–the world’s ugliest color will be helpful in smoking rates in their country, according to a team of experts ... researchers found that pantone publications, including pantone 448c visually, are chock full of “ ugly ” reactions, including ...
DocLing	... world’s ugliest color - will be stripped of colored branding and replaced with a sludge-like color .in may, previously passed legislation will go into effect requiring all packs of cigarettes to be standardized ...
Ours	–the world’s ugliest color is serving an important purpose: ... more likely to deter smoking from reaching for their next pack of cigarettes ... it will go into effect requiring to be standardized ... found that smokers aged 15 years and over half a year ... in smoking rates than those in 2002 and 48.6 % in 2002 ...

TABLE V: Example of source documents and summaries.

Objectives	R-1	R-2	R-SU
SSL	44.64	15.47	17.83
TL	44.15	14.98	17.52
SL	44.10	15.00	17.47
CPL	45.16	15.39	18.48
CPL-R	45.03	15.09	18.40
DPL	44.74	15.33	18.05
DPL-N	44.41	14.93	17.86

TABLE VI: Models performance with different objective functions on Multi-News validation dataset. “R” and “N” indicates randomly sort documents in the same document set and normalization.

D. Objective Function Selections

We evaluate model effectiveness by integrating shared and specific representations using various objective functions. This evaluation aims to highlight capturing common and specific information in MDS and understand how different training objectives affect model performance.

1) *Specific-Shared Loss (SSL) V.S. Triplet Loss (TL)*: Inspired by the good performance of specific representations, we intuitively think that disentangling the shared representation may further improve the model performance. Here shared representation refers to common information shared by multiple documents in one set. To obtain the shared representations, we incorporate a shared encoder $\omega(\cdot; \nu)$ to learn the shared representations \mathbf{H}_i from the document \mathbf{d}_i by:

$$\mathbf{H}_i = \omega(\mathbf{d}_i; \nu), \quad (10)$$

where ν is a set of learnable parameters. Within a set, shared representations should be similar, while specific representations should be distinguishable. We use two objective functions to achieve this: one to encourage similarity in shared representations and another to emphasize distinctiveness in specific representations. Specific-shared loss represents by:

$$L_{total} = L_{gen} + \beta \cdot L_{spec} + \gamma \cdot L_{shared}, \quad (11)$$

$$L_{shared} = \sum_{i=1}^N L_{shared}^i, \quad (12)$$

$$L_{shared}^i = \begin{cases} \|\mathbf{H}_i - \mathbf{H}_{i+1}\|_p & i \neq N \\ \|\mathbf{H}_N - \mathbf{H}_1\|_p & i = N \end{cases}, \quad (13)$$

$$L_{spec} = \sum_{i=1}^N \left\| \mathbf{S}_i^\top \mathbf{H}_i \right\|_F^2, \quad (14)$$

where γ is a balance factor. For a document d_i , we expect its specific representations to be orthogonal with its shared representations.

Triplet loss is inspired by contrastive representation learning from positive and negative samples [28]. For a document d_i , its shared and specific representation, \mathbf{H}_i and \mathbf{S}_i , can be seen as an anchor and a negative sample. The positive sample can be another shared representation map \mathbf{H}_j from document d_j . Note that d_i and d_j are in the same document set. The final objective function can be:

$$L_{total} = L_{gen} + \beta \cdot L_{triplet}, \quad (15)$$

$$L_{triplet} = \sum_{i=1, j \neq i}^N \max(\|\mathbf{H}_i - \mathbf{H}_j\|_2^2 + \|\mathbf{H}_i - \mathbf{S}_i\|_2^2), \quad (16)$$

Table VI reveals the performance of the SSL-equipped model surpasses that of the TL-equipped one but does not surpass the results achieved by the Disentangle model’s objective. We hypothesize the conflict between shared and document-set representations causes this phenomenon during the optimization of summary generation. As a result, we decide not to incorporate the shared and specific representations together.

2) *Specific Loss V.S. Shared Loss (SL)*: Given the better performance of specific-shared loss than triplet loss, we further examine the specific representations and the shared representations separately. We compare the results of specific loss and shared loss. The total objective function to generate shared representations can be defined as:

$$L_{total} = L_{gen} + \gamma \cdot L_{shared}, \quad (17)$$

where the calculation of L_{shared} is equal to Equation (12). In Table VI shows that the model equipped with shared loss obtains lower performance than that equipped with specific loss. Furthermore, to dig out why the specific loss has a comparative advantage in summaries generation, we visualize the attention maps (Figure 2) of specific and shared representations from the last encoding layer. Interestingly, the attention-specific encoder is mainly focused on the individual words “hue”

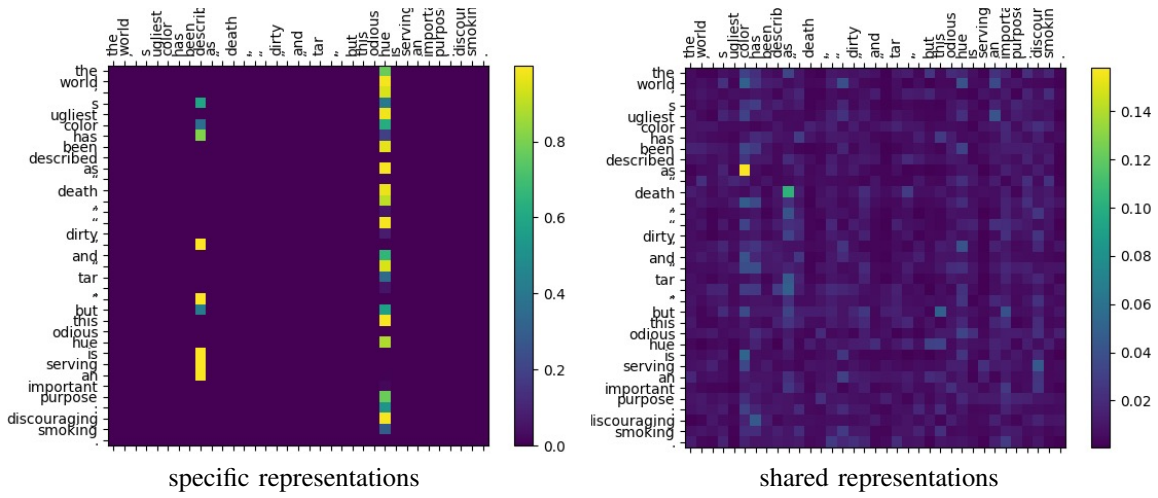


Fig. 2: Attention maps of learned specific representations and shared representations.

Variants	R-1	R-2	R-SU	Variants	R-1	R-2	R-SU
$\alpha = 1$	43.60	13.94	17.30	$\beta = 1$	44.15	15.17	17.72
$\alpha = 0.1$	43.36	13.67	17.26	$\beta = 0.1$	44.24	15.16	17.77
$\alpha = 0.01$	45.16	15.39	18.48	$\beta = 0.01$	44.64	15.45	18.09
$\alpha = 0.001$	44.67	15.50	17.90	$\beta = 0.001$	45.16	15.39	18.48
w/o Spec Feat	43.66	14.79	17.39	w/o Spec Loss	43.66	14.79	17.39

TABLE VII: Model performance on Multi-News validation set by tuning specific representation trade-off factor α and loss trade-off factor β .

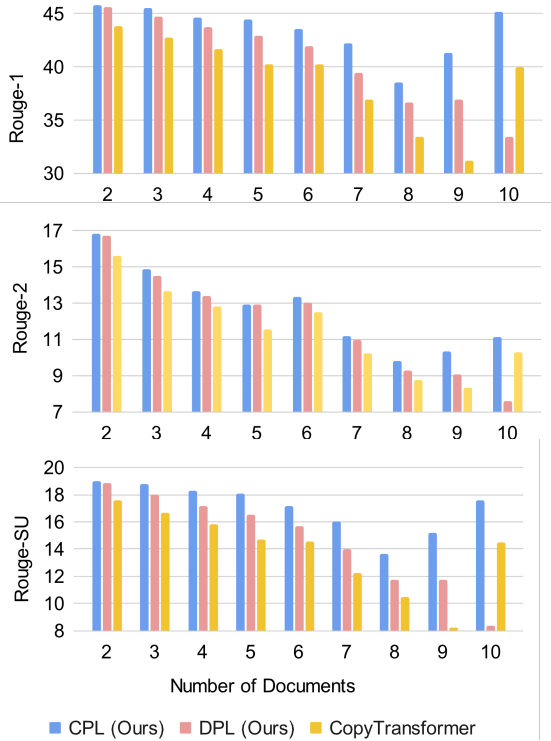


Fig. 3: ROUGE scores of DisentangleSum with circle-paired-loss (CPL) and dense-paired-loss (DPL), and CopyTransformer on document sets containing two to ten documents.

which is the specific information for #1 document. However, the heatmap of shared representations is more scattered than the specific representations. This may be because specific representations concentrate on important information of each document while shared representations do not. Consequently, we opt not to select the objective function associated with shared representations for our main experiment.

3) *The Selection of Specific Loss*: This section investigates two design options for specific loss: circle-paired loss (CPL), introduced in section III-C, and dense-paired loss (DPL). DPL involves computing specific representation loss for each pair of documents in the same document set. We conduct two experiments in this subsection:

(1) **Compare the overall performance of DisentangleSum equipped with CPL and DPL.** We evaluate the models on the Multi-News dataset, analyzing their performance with different objectives. The results in Table VI indicate that CPL outperforms DPL across all three evaluation metrics.

(2) **Explore the impact of document number on specific loss.** To investigate the relationship between specific loss and the number of documents, we divided the Multi-News validation set into subsets based on the document set size. We compare the model performance trained with CPL and DPL on these subsets. Figure 3 illustrates that DisentangleSum with CPL outperforms DPL and CopyTransformer across all subsets in terms of three ROUGE scores. Notably, when the document set size is two, the results of DPL and CPL are quite similar for the ROUGE-1 score. However, as the number of

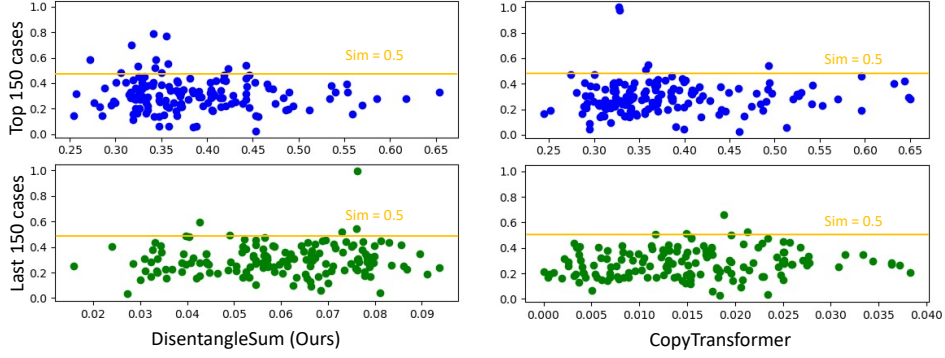


Fig. 4: The distribution of document similarity scores in the Top 150 and Last 150 cases. The X-axis and Y-axis of each sub-figure are ROUGE-SU scores (scale to 0 ~ 1) and documents similarity scores, respectively. The orange line represents the document similarity score equal to 0.5.

documents increases, the model trained with DPL experiences a significant performance drop, while the model trained with CPL exhibits a slower decline. This trend holds for ROUGE-2 and ROUGE-SU scores as well. Besides, from the perspective of computational complexity, as the number of documents increases, the document pairs in DPL increase quadratically (e.g. 10 documents yield 45 pairs), while CPL does not.

In order to exclude the impacts of the order of documents in a document set and the loss scale, we further conduct two experiments: (1) Based on CPL, we randomly sort documents in the same document set; (2) Based on DPL, we adjust the loss scale through normalization, dividing the right hand side of Equation 5 by N^2 . The results (in Table VI) ruled out the interference of these two factors. The performance differences may be that for MDS tasks, the documents in the same document set describe topic-relevant concepts, yet with some document-specific information. The constraint may be too strong by imposing a model to learn document-specific representations completely different between documents, which in turn may incur a confused model and less “informative” representations learned.

E. Hyperparameter Scale of Models

We perform a hyperparameter study to examine the effectiveness of specific representation trade-off factor α and loss trade-off factor β , controlling the trade-off strength of fetching the document specific information and document-set information. The results are shown in Table VII. Both the weights of α and β are controlled by searching the grid [1, 0.1, 0.01, 0.001, 0]. The experiments of evaluation α is performed under β equals to 0.001; while the examination β is conducted by setting α to 0.01. By setting either specific representation weights or specific loss weights to 0s, the model performance is significantly degraded. It suggests the positive contribution of grasping documentary unique information. Interestingly, with the increasing of specific representation trade-off factor α from 0.001 to 0.01, the ROUGE scores generally have an increasing trend. But the score goes up and then goes down

when α is from 0.01 to 1. The optimal choice of the hyperparameter α falls in the middle of the evaluated values, which is 0.01. Similar results show for the experiments of loss trade-off factor β . Generally, 0.001 is recommended for β to achieve the best performance. The experimental results indicate that the existence of the document-specific representation learner and the orthogonal constraint of document-specific representation generation is important. Meanwhile, setting large α and β obstructs model optimization and summary generation.

F. DisentangleSum Performances with Different Inter-document Similarities

The aim of this subsection is to analyze the correlation between DisentangleSum performance and inter-document similarities. We define a simple function to calculate the document similarity within a document set using statistical analysis:

$$Sim(\mathbf{D}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{2 \cdot \text{overlap}(\mathbf{d}_i, \mathbf{d}_j)}{N(N-1)}, \quad (18)$$

where N represents the document number in each document set. It calculates the content overlap between each pair of documents within the document set. We evaluate the DisentangleSum and CopyTransformer models by calculating ROUGE scores for document sets and ranking them. We analyze the Top 150 and Last 150 cases, finding average similarity scores of 0.308 and 0.299 for DisentangleSum, and 0.293 and 0.281 for CopyTransformer. Figure 4 shows: (1) DisentangleSum’s Top 150 cases have slightly higher document similarity scores than the Last 150 cases. (2) DisentangleSum’s Top 150 cases have more instances with similarity scores above 0.5 compared to the Last 150 cases and CopyTransformer’s Top 150 cases.

These findings suggest that the proposed model tends to perform better when the document similarity score is higher. The potential reason is in one document set when the overlap between each document is relatively large, the ratio of the uniqueness of each document is relatively small. Models that do not explicitly capture document-specific information may struggle to capture the specific details from the source

documents. The DisentangleSum model, designing to retain document-specific information, performs better in such cases when the document similarity score is higher.

V. CONCLUSION

In this paper, we introduce DisentangleSum, a framework to disentangle document-specificity for better abstractive multi-document summarization representations. To optimize the specific representation learning, we apply an orthogonal constraint to encourage the document-specific representation learner to catch specific information per document. The experiments on two prevalent datasets show the superior performances of the proposed model over other counterparts. Furthermore, we also provide extensive analyses that reveal DisentangleSum exhibits broader coverage of input documents and better preservation of document-related information. These analyses help researchers understand the intuitiveness of the proposed model and could serve as an informative reference to the multi-document summarization research community.

ACKNOWLEDGMENTS

This research is partially supported by Australian Research Council (ARC) Discovery Project DP230100233.

REFERENCES

- [1] C. Ma, W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng, "Multi-document summarization via deep learning techniques: A survey," *ACM Computing Surveys (CSUR)*, 2022.
- [2] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, "Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model," in *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, pp. 1074–1084.
- [3] Y. Mao, Y. Qu, Y. Xie, X. Ren, and J. Han, "Multi-document summarization with maximal marginal relevance-guided reinforcement learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, Online, pp. 1737–1751.
- [4] C. Zhao, T. Huang, S. B. R. Chowdhury, M. K. Chandrasekaran, K. R. McKeown, and S. Chaturvedi, "Read top news first: A document re-ordering approach for multi-document news summarization," in *Findings of the Association for Computational Linguistics (ACL 2022 findings)*, Dublin, Ireland, pp. 613–621.
- [5] Y. Liu and M. Lapata, "Hierarchical transformers for multi-document summarization," in *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, pp. 5070–5081.
- [6] W. Li, X. Xiao, J. Liu, H. Wu, H. Wang, and J. Du, "Leveraging graph to improve abstractive multi-document summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online, pp. 6232–6243.
- [7] H. Jin, T. Wang, and X. Wan, "Multi-granularity interaction network for extractive and abstractive multi-document summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online, pp. 6244–6254.
- [8] Y. Song, Y. Chen, and H. Shuai, "Improving multi-document summarization through referenced flexible extraction with credit-awareness," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022)*, Seattle, United States, pp. 1667–1681.
- [9] R. Pasunuru, M. Liu, M. Bansal, S. Ravi, and M. Dreyer, "Efficiently summarizing text and graph encodings of multi-document clusters," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2021)*, Online, pp. 4768–4779.
- [10] D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang, "Heterogeneous graph neural networks for extractive document summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online, pp. 6209–6219.
- [11] R. Wolhandler, A. Cattani, O. Ernst, and I. Dagan, "How "multi" is multi-document summarization?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, Abu Dhabi, United Arab Emirates, pp. 5761–5769.
- [12] X. Wan and J. Yang, "Multi-document summarization using cluster-based link analysis," in *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, pp. 299–306.
- [13] O. Ernst, A. Caciularu, O. Shapira, R. Pasunuru, M. Bansal, J. Goldberger, and I. Dagan, "Proposition-level clustering for multi-document summarization," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022)*, Seattle, United States, pp. 1765–1779.
- [14] J. Zhao, M. Liu, L. Gao, Y. Jin, L. Du, H. Zhao, H. Zhang, and G. Haffari, "Summpip: Unsupervised multi-document summarization with sentence graph compression," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, Online, pp. 1949–1952.
- [15] C. Ma, W. E. Zhang, P. D. D. Pitawela, Y. Qu, H. Zhuang, and H. Wang, "Document-aware positional encoding and linguistic-guided encoding for abstractive multi-document summarization," in *Proceedings of the IEEE Word Congress on Computational Intelligence (WCCI 2022)*, Padua, Italy.
- [16] W. Xiao, I. Beltagy, G. Carenini, and A. Cohan, "PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, Dublin, Ireland, pp. 5245–5263.
- [17] Y. Lu, Y. Dong, and L. Charlin, "Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, Online, pp. 8068–8074.
- [18] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Workshop of Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.
- [19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, Addis Ababa, Ethiopia.
- [20] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2018)*, New Orleans, United States, pp. 708–719.
- [21] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," vol. 22, 2004, pp. 457–479.
- [22] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain, pp. 404–411.
- [23] J. G. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (SIGIR 1998)*, Melbourne, Australia, pp. 335–336.
- [24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, USA, pp. 5998–6008.
- [25] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, Canada, pp. 1073–1083.
- [26] S. Xu, H. Li, P. Yuan, Y. Wu, X. He, and B. Zhou, "Self-attention guided copy mechanism for abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online, pp. 1355–1362.
- [27] S. Liu, J. Cao, R. Yang, and Z. Wen, "Highlight-transformer: Leveraging key phrase aware attention to improve abstractive multi-document summarization," in *Findings of the Association for Computational Linguistics (ACL/IJCNLP 2021)*, Online, pp. 5021–5027.
- [28] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston, USA, pp. 815–823.