SALAS: Supervised Aspect Learning Improves Abstractive Multi-document Summarization Through Aspect Information Loss

Haotian Chen(✉), Han Zhang, Houjing Guo, Shuchang Yi, Bingsheng Chen, and Xiangdong Zhou

School of Computer Science, Fudan University, Shanghai, China
{htchen18,hanzhang20,xdzhou}@fudan.edu.cn,
{houjingguo21,scyi21,chenbs21}@m.fudan.edu.cn

Abstract. Abstractive multi-document summarization (MDS) aims at summarizing and paraphrasing the salient key information in multiple documents. For dealing with the long-input issue brought by multiple documents, most previous work extracts salient sentence-level information from the input documents and then performs summarizing on the extracted information. However, the aspects of documents are neglected. The limited ability to discover the content on certain aspects hampers the key information seeking and ruins the comprehensiveness of the generated summaries. To solve the issue, we propose a novel **S**upervised **A**spect-**L**earning **A**bstractive **S**ummarization framework (SALAS) and a new aspect information loss (AILoss) to learn aspect information to supervise the generating process heuristically. Specifically, SALAS adopts three probes to capture aspect information as both constraints of the objective function and supplement information to be expressed in the representations. Aspect information is explicitly discovered and exploited to facilitate generating comprehensive summaries by AILoss. We conduct extensive experiments on three public datasets. The experimental results demonstrate that SALAS outperforms previous state-of-the-art (SOTA) baselines, achieving a new SOTA performance on the three MDS datasets. We make our code for SALAS publicly available (https://github.com/Hytn/AspectSum).

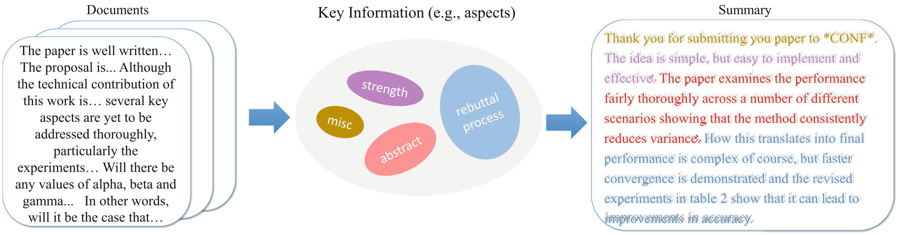
Keywords: Multi-document summarization · Supervised aspect learning · Aspect information loss

1 Introduction

Document summarization (DS) aims to convert a document or multiple thematically related documents into a fluent, condensed, and informative summary [11,18,23]. It is beneficial for a wide range of downstream applications including generating Wikipedia abstracts [15,30], creating news digests [1], and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
D. Koutra et al. (Eds.): ECML PKDD 2023, LNAI 14172, pp. 55–70, 2023.
https://doi.org/10.1007/978-3-031-43421-1_4

opinion summarization [2]. The given documents comprise various aspects and can overlap and complement each other [11]. Therefore DS faces more challenges due to capturing and organizing information scattered across the long input (especially entailed by multiple documents) [7,18].



The paper is well written… The proposal is... Although the technical contribution of this work is… several key aspects are yet to be addressed thoroughly, particularly the experiments… Will there be any values of alpha, beta and gamma... In other words, will it be the case that…

Thank you for submitting you paper to *CONF*. The idea is simple, but easy to implement and effective. The paper examines the performance fairly thoroughly across a number of different scenarios showing that the method consistently reduces variance. How this translates into final performance is complex of course, but faster convergence is demonstrated and the revised experiments in table 2 show that it can lead to improvements in accuracy.

Fig. 1. An example for performing multi-document summarization by aspect information.

Traditional methods tackled DS based on feature engineering [7], statistical learning [3], and graph theory [7,20]. Most of them extract salient textual units, structural dependencies among phrases, keywords, or semantic clusters as key information to aid the generation of a final summary [18]. Recently, pre-trained language models (PLMs) significantly facilitate MDS through several paradigms, including adopting extract-then-generate methods [26,28] and applying hierarchical models architecture [15,29]. The former shorten the length of input context by extracting salient texts while the latter improve the capability of models to simultaneously process all information. It is reported that extract-then-generate methods, hierarchical models, most traditional methods, and human performance possess a common background: They believe that a summary can be produced through a top-down method where key information is first detected from the input documents explicitly and then summarized or adopted to guide the summarization [12,19].

However, aspect information is rarely considered or modeled as the key information in these methods. Summaries are often viewed as plain text despite the fact that their summarized documents can be well organized and written down according to the underlying aspects (e.g., writing according to a mind map). As shown in Fig. 1, the detected aspects, as key information, can summarize the input documents and thereby guides the top-down methods to generate a final summary. A pipeline method [30] is proposed to address the issue, which first detects topics described by input documents, and then encodes the detected topics together with the documents to generate a final summary. Despite the improved performance, the method leaves two problems unsolved: (1) the separately-trained pipeline methods can suffer from cascade errors; (2) the aspect information indirectly aids the generator in an implicit way. We argue that aspect information is essential for supervising the generation process of MDS.

In this paper, we propose a novel supervised aspect learning abstractive summarization framework (SALAS) with aspect information loss (AILoss). AILoss enables SALAS to capture aspect information as both constraints of the objective function and sufficient expressive power of representations to guide the generating process. Specifically, we design three linear probes to detect the aspect information expressed by the representations of both input documents and the generated summaries: an encoder probe for documents and two decoder probes for summaries. AILoss considers the detected aspect information from three probes, which not only infuses representations with aspect information but also eliminates the inconsistency between the aspects expressed by documents and summaries. It renders summaries to cover each aspect mentioned in corresponding documents. The aspects and summaries are jointly learned with our proposed AILoss. We evaluate our proposed SALAS on 3 MDS benchmarks. Using the same backbones, SALAS outperforms the strong baseline models and achieves a new state-of-the-art performance on the three benchmarks. Our main contributions are threefold:

- We introduce SALAS, a novel aspect-guided joint learning summarization framework that captures aspect information to guide the generating process.
- We propose aspect information loss (AILoss) to constrain the objective function and give sufficient expressive power to representations, which aids the generating process.
- Experimental results show that SALAS significantly outperforms previous SOTA methods on 3 MDS benchmarks. We further conduct a comprehensive analysis that exhibits the high quality of detected aspects and the effectiveness of modules in SALAS.

2 Related Work

2.1 Multi-document Summarization

Traditional MDS obtain key information represented by words, sentences, graphs, and semantic clusters to guide the generator [7,20]. With recent significant improvement in SDS brought by large-scale PLMs [13], most researchers tackle MDS based on PLMs in four ways with two underlying purposes: (1) To enhance the long-input processing capability of models, they propose sparse attention [4] and hierarchical model architectures [15,29]. The former is proposed for reducing the memory complexity of transformer-based PLMs while the latter is designed for capturing dependency information among sentences and words. (2) To shorten the length of source input, researchers adopt extract-then-generate methods [26,28] and divide-and-conquer approaches [8]. The former extracts salient texts (key information) from the given documents and then summarizes them, the latter divides the given documents into sections and then individually summarizes them to form a final summary.

The extract-then-generate methods and hierarchical model architectures try to collect and merge the information scattered across the source input in a heuristic way and then summarize the derived key information. Inspired by the paradigm, we focus on exploring and modeling the aspect information, which is objective, definite, concise, and often neglected in previous work, as key information to explicitly supervise the generating process of summaries.

2.2 Aspect-Related Text Generation

Little recent work exploits aspect information while generating a generic summary [21]. One line of research focuses on heuristically identifying aspects (e.g., words or phrases) expressed in opinions for opinion summarization and sentiment analysis [25]. Aspect information and its corresponding context will help distinguish the sentiment polarities of reviews about different aspects of a product. A previous work [1] also proposes a summarization system where aspect-level keywords can be automatically extracted without assuming human-annotated data for training the extractor.

More recently, aspect information is manually annotated in some aspect-oriented abstractive summarization datasets in the multi-document setting. It includes WikiAsp for aspect-oriented Wikipedia summarization [9], summaries of popular and aspect-specific customer experience (SPACE) dataset for opinion mining [2], and meta-review dataset (MRED) for structure-controllable meta-review generation [24], which makes human-annotated aspect information sufficient and available for text generators.

Instead of directly applying the heuristically extracted aspect information with noises, we propose modeling the human-annotated accurate aspect information to facilitate abstractive MDS. TWAG [30] explores a pipeline method to model the aspect information for abstractive MDS, which we compare with.

3 Methodology

We present the overview of our framework composed of two kinds of aspect probes and a generator with aspect constraints in Fig. 2 and then elaborate on each component in the following sections. In Sect. 3.1, we first formulate the target task and our proposed aspect-guided framework. In Sect. 3.2 and Sect. 3.3, we then introduce two kinds of aspect probes for documents and summaries, respectively. After that, we elaborate on the mechanism of our proposed aspect-guided generator in Sect. 3.4. Finally, we summarize and formulate the training objective in Sect. 3.5.

3.1 Task and Framework Formulation

We formulate the task and our proposed solution and then introduce the underlying motivation in this section. In the MDS task, the input document set $\mathcal{D} = \{D_i\}_{i=1}^{n}$ comprises multiple documents and can be expressed by its concatenate context X, while the generated output is their summary y of length T. Given input documents X and the previously generated tokens $y_{<t}$, the goal

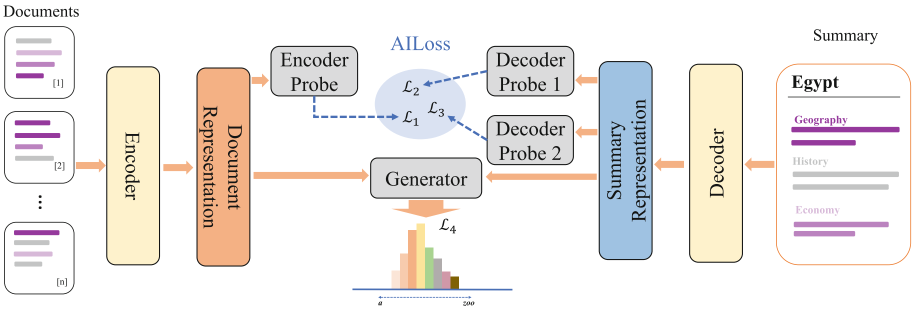


Fig. 2. An overview of our framework.

of previous methods in MDS is to train a learning model, which aims to maximize the likelihood of the given optimal summary y^*, to generate a sequence of summary tokens y which can be described by,

$$y^* = \arg\max_y P(y \mid \mathcal{D}) = \arg\max_y \prod_{t=1}^{T} P\left(y_t \mid \mathcal{D}, y_{<t}\right). \tag{1}$$

However, most documents are well organized and written down according to the underlying aspect information, which guides human-written summaries of multiple documents [30]. That is to say, aspect information is a constraint of the summary-generating process. The constraint is missed in Eq. (1) adopted by previous work, leading to their increased risk of falling into suboptimal results. To solve the issue, we present two kinds of probabilistic models (probes) $P(\mathcal{A} \mid \mathcal{D})$ and $P(\mathcal{A} \mid y)$ to estimate the aspect information $\mathcal{A} = \{a_i\}_{i=1}^{N}$ contained in the representations of input documents and generated words. In each step t of the generating process, the imbalance in the amount of two source aspect information supervises the generator to recover the missing target aspect information. To measure the extent to which aspect information is missed, we introduce aspect information loss (AILoss). Let d be a symmetric "distance" of the amount of aspect information contained by two representations, where d possesses a variety of choices, including L_2 distance and symmetric KL-divergence. The AILoss is defined as follows:

Definition 1 (Aspect Information Loss). *The aspect information loss between the aspect information represented by source input text C and the target amount of aspect information P^* is*

$$\rho = d\left[P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid C\right), P_i^*\right]. \tag{2}$$

The representation of source input text is ξ-informative if $\rho \le \xi$.

Note that in some cases, we have restricted access to the exact number of categories of aspects and to the human-annotated aspect information for context, which exacerbates the difficulty of estimating aspect information. Thanks to methods for aspect-level keywords extraction [1] and clustering-based unsupervised representation learning models [10,27], models are able to extract and classify aspect information in a heuristic way. However, noise information will be introduced during the process to confuse the generator. As it remains unclear whether aspect information can facilitate the generation of generic summaries in MDS and which way can effectively exploit the aspect information. In this paper, we demonstrate the effectiveness of both aspect information and our framework in a convincing and clear setting and leave the situation where aspect information is combined with noise to be studied in future work.

3.2 Encoder Probe

The encoder probe aims to detect the aspect information contained in representations of the input documents. For input context X and the previously generated words $y_{<t}$, PLMs generate both token and context representations. We obtain the context representations of X and $y_{<t}$ by,

$$\mathbf{h}_{\text{doc}} = \text{PLM}(X), \mathbf{h}_{\text{sum}} = \text{PLM}(y_{<t}), \tag{3}$$

where PLM includes multiple options, such as BERT [6], BART [13], GloVe [22], etc., exhibiting different features explained by representations. To give sufficient expressive power to the context representations, we integrate aspect information into the representations by training them together with an aspect probe,

$$\mathcal{L}_1 = \sum_{a_i \in A} d\Big[P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid X; \phi\right), P_{\text{gold}}(a_i)\Big], \tag{4}$$

where d denotes the distance between the prediction and its corresponding ground truth $P_{\text{gold}}(a_i)$. A represents the set of target aspects that compose the input documents and their corresponding summary. We adopt the commonly used mean square error (MSE) as our loss function to compute the overall distance $\mathcal{L}_1$. Aspect probe $P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid X; \phi\right)$ with trainable parameters ϕ maps the input documents to a certain score of each aspect and then derives the corresponding output by the sigmoid function σ,

$$P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid X; \phi\right) = \sigma\left(\mathbf{W}_i \mathbf{h}_{\text{doc}} + \mathbf{b}_i\right), \tag{5}$$

where ϕ consists of $\mathbf{W}$ and $\mathbf{b}$. The output can either be regarded as probability or measurement, which indicates the amount of a certain kind of aspect information existing in the input context.

3.3 Decoder Probe

The decoder probe aims to detect the aspect information contained in representations of generated words. We use a similar method to derive aspect information from the previously generated words. Specifically, during the process of generating, the amount of aspect information dynamically changes. If we discard words from the ground truth summary one after another, the corresponding aspect information becomes less and less. Given the target amount of aspect information $P(a_i)$ and the generated words $y_{<t}$ with length t, where the ground truth words summarize the current aspect a_i with length L_i and other words describe other aspects with length L_i^-, we give out a function $F_i(t)$ to measure the expected amount of increasing information for aspect a_i in the t step (t-th newly generated word). We assume that each word equally contributes to the information of each aspect, then $F_i(t)$ can be represented by,

$$F_i(t) = \frac{t - L_i^-}{L_i} P(a_i). \tag{6}$$

Note that the target amount of information $P(a_i)$ can either be ground truth probability (derived by human-annotated aspect labels or introduced from aspects possessed by input documents) which equals 1 or be probability predicted by the aspect probe for input documents. Correspondingly, we propose two decoder probes as follows.

On human-annotated datasets where aspect information is accurate and available, we adopt ground truth aspect information in training where $P(a_i) = 1$. Similar to the motivation and training process of the encoder probe for input documents, we train the first decoder probe $P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid y_{<t}; \eta\right)$ and representations of generated words by,

$$\mathcal{L}_2 = \sum_{t=1}^{T} \sum_{a_i \in A} d\Big[P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid y_{<t}; \eta\right), \frac{t - L_i^-}{L_i}\Big]. \tag{7}$$

During the training process, the probed scores for aspects represent the comprehensiveness of the generated summary and serve as an indicator to discover the missing aspect information in the previously generated words.

We can apply the second decoder probe on any datasets, including those where we have restricted access to aspect labels. It infuses the probed aspect information $P(a_i) = P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid X; \phi\right)$ from input documents into representations of the generated words,

$$\mathcal{L}_3 = \sum_{t=1}^{T} \sum_{a_i \in A} d\Big[P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid y_{<t}; \mu\right), \frac{t - L_i^-}{L_i} P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid X; \phi\right)\Big]. \tag{8}$$

Here, we do not propagate gradients of $\mathcal{L}_3$ to the parameters ϕ to avoid introducing spurious correlations which can impede the learning process described in Eq. (4).

3.4 Aspect-Guided Generator

As we obtain the gap between aspect information in input documents and that in the generated words, we require a generator, which abides by the regularity described in Eq. (6) where the gap is continuously narrowed during the generating process, to recover the missing aspect information when generating the summary. We exhibit two reasons why a generator learns to generate those words that can narrow the gap: (1) The ground truth aspect information in documents is instructive and easily accessible as mentioned in Sect. 3.1. Besides the precious human-annotated aspect information for summaries, it can serve as a supplementary indicator to supervise the generation of summaries. (2) The gap reveals missing aspects. The existence of a gap indicates that salient information of a certain aspect from input documents does not occur in the summary. The learning model minimizes the overall loss to narrow the gap, which constrains the generating process and thus reduces the entropy of the potential words. Therefore, the generation loss for a generated summary is represented by,

$$\mathcal{L}_4 = -\frac{1}{T}\sum_{t=1}^{T} \log P_\Theta\left(y_t \mid y_{<t}, X\right), \tag{9}$$

where Θ denotes model parameters excluding the parameters of three probes.

3.5 Training Objective

The core idea of our proposed learning framework is to train a summarization model under the constraints of aspect information abstracted from the input documents, which not only supervises the generation, but also avoids the isolation between the aspect probes and the generator. Different from the methods that exploit and incorporate aspect information by,

$$y^* = \arg\max_{\mathcal{A}} P(\mathcal{A} \mid \mathcal{D}) \arg\max_{y} P(y \mid \mathcal{D}, \mathcal{A}), \tag{10}$$

which can suffer from the cascade error and impede the interaction between the aspect probe and summarization model. Our proposed training objective is represented by,

$$\arg\max_{\Theta} \prod_{t=1}^{T} P_\Theta\left(y_t \mid X, y_{<t}\right)$$

$$\text{s.t.} \sum_{t=1}^{T}\sum_{a_i \in A} d\left[P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid y_{<t}\right), F_i(t)\right] + \sum_{a_i \in A} d\Big[P_{\mathcal{A}|\mathcal{D}}\left(a_i \mid X\right), P_{\text{gold}}(a_i)\Big] \leq \xi, \tag{11}$$

where $i = 0, 1, \ldots, N$. $P(a_i)$ in $F_i(t)$ denotes either ground truth or the probability of a_i probed from X and ξ an upper bound of inconsistency between the target amount of aspect information and that probed from documents and generated words.

To sum up, the overall training objective of our proposed framework is

$$\mathcal{L} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 + \lambda_4\mathcal{L}_4, \tag{12}$$

where λ_1, λ_2, λ_3, and λ_4 are hyperparameters to balance and control the influence of different loss components. Parameters ϕ are solely optimized with $\mathcal{L}_1$, parameters η are solely optimized with $\mathcal{L}_2$ and $\mathcal{L}_3$, and parameters Θ are optimized with $\mathcal{L}$.

4 Experiments

4.1 Datasets

We experiment on two datasets as follows.
MRED [24] is a highly abstractive dataset focusing on meta-reviews from a peer-reviewing system (ICLR) which contains essential and high-density opinions. It is provided for structure-controllable text generation. The meta-reviews are manually written to summarize the aspects described in different reviews and we use *sent-ctrl* version of MRED.
WikiAsp [9] is provided for aspect-based summarization. It contains articles, consisting of section titles and section texts, from various domains of Wikipedia and their corresponding reference documents. The section texts serve as aspect-based summaries of corresponding reference documents. We randomly select 2 out of 20 domain datasets from WikiAsp as evaluation benchmarks. The two datasets are **Historic Place** and **Plant**.

4.2 Baselines

We compare SALAS with previous state-of-the-art methods (comprising an extractive model and 4 abstractive models) on the three datasets:
TextRank [20] is a common extractive summarization baseline model which uses vertex scores calculated by a graph-based "random-surfer model" to rank sentences.
TWAG [30] is a two-step abstractive summarization method that first detects the aspects described by the multiple source documents and then performs summarization based on the detected aspects.
BertAbs [16] is an abstractive summarization model with encoder initialized with BERT [5] and transformer decoder randomly initialized.
Longformer [4] is a pre-trained language model tackling long input by sparse attention. Following BertAbs [16], We initialize the encoder with Longformer and randomly initialize the transformer decoder.
BART [13] is a SOTA abstractive summarization model pre-trained with the objective of denoising autoencoding.

We also compare with other baselines mentioned in the work proposing the corresponding dataset.

4.3 Implementation Details

We complete our experiments on a single RTX3090 GPU. We first load the pre-trained models released by Huggingface[1] as the backbones. To keep in line with the basic settings of baselines for fair comparison, we adopt the common hyper-parameters used in the transformer-based baseline models. Specifically, we apply AdamW algorithm [17] to optimize model parameters with a learning rate of 1e-5. We evaluate our generated summaries against the reference manually written

[1] https://huggingface.co/models.

ones by calculating the F_1-scores of ROUGE_1, ROUGE_2, and ROUGE_L [14]. Following previous work, we adopt the Rouge evaluation script[2] provided by Huggingface with "use_stemmer" enabled.

4.4 Main Results

We compare SALAS with all of the previous SOTA methods on MReD and WikiAsp. We also further implement several common strong baselines for comparison and deeper analysis. Table 1 and Table 2 show the main results of the baseline models and our proposed SALAS. We can observe that SALAS outperforms the existing SOTA baselines on MReD and Wikiasp in BERT, Longformer, and BART backbones, respectively. Specifically, with $\text{BERT}_{\text{base}}$ and $\text{BART}_{\text{large}}$ as the PLM, SALAS surpasses BertAbs and $\text{BART}_{\text{large}}$ by 2.02/0.50/0.48 and 0.58/2.99/0.79 of ROUGE-1/2/L scores respectively, achieving new SOTA performance on MReD. Meanwhile, SALAS yields gains of 2.13/0.69/0.84 and 1.44/0.83/0.84 of ROUGE-1/2/L scores on two randomly selected domains from Wikiasp compared to the previous SOTA methods. The experimental results show the effectiveness of the overall framework of SALAS.

Table 1. Performance on MReD. The signal † denotes that the results of models are quoted in the original paper proposing MReD. The rest of the results are based on our implementation.

Model	R-1	R-2	R-L
MMR†	32.37	6.28	17.58
LexRank†	32.60	6.66	17.48
TextRank†	33.52	7.20	17.75
TWAG	27.82	7.22	19.99
Longformer	23.39	5.63	20.52
BertAbs-$\text{BERT}_{\text{base}}$	23.57	6.67	18.59
SALAS-$\text{BERT}_{\text{base}}$	25.59	7.17	19.07
$\text{BART}^{\dagger}_{\text{large}}$	38.59	10.61	22.93
SALAS-$\text{BART}_{\text{large}}$	**39.17**	**13.60**	**23.72**

We attribute the improvement to the incorporation of aspect information and our proposed joint learning framework for two reasons. First, aspect information significantly improves the performance of models. We observe the performance gaps (the aforementioned gains of ROUGE-1/2/L scores) between models that adopt SALAS framework and models with the same backbones neglecting aspect information. The difference between two kinds of models is that the former incorporates aspect information by constraining the objective function, which indicates that by properly infusing models and constraining the objective function

[2] https://github.com/huggingface/transformers/blob/main/examples/pytorch/summarization/.

with aspect information, models are able to achieve more significant improvements. Second, learning aspect information in a joint way largely enhances the effectiveness of models. Compared to TWAG which models aspect information in a two-step way, our proposed SALAS significantly outperforms TWAG by 11.35/6.38/3.73, 10.17/4.12/2.87, and 6.11/2.23/1.51 ROUGE-1/2/L on MReD, Historic Place domain of WikiAsp, and Plant domain of WikiAsp, respectively.

Table 2. Performance on Wikiasp. All of the results are based on our implementation.

Model	Historic Place			Plant		
	R-1	R-2	R-L	R-1	R-2	R-L
TextRank	10.97	2.05	7.50	14.12	1.95	8.70
TWAG	24.36	7.07	16.98	22.51	4.68	14.99
BERT$_{\text{base}}$	25.22	8.21	14.88	21.84	6.35	13.54
Longformer	33.17	10.38	18.82	25.67	6.21	15.67
SALAS-Longformer	34.30	10.65	19.39	25.77	6.02	15.76
BART$_{\text{large}}$	32.40	10.50	19.01	27.18	6.08	15.66
SALAS-BART$_{\text{large}}$	**34.53**	**11.19**	**19.85**	**28.62**	**6.91**	**16.50**

4.5 Results of Ablation Study

We conduct ablation experiments on SALAS to further test the effectiveness of its components. As shown in Table 3, we observe that each component exerts a positive effect on the performance of SALAS on all of the three datasets, demonstrating their effectiveness. Furthermore, the extents of their influence differ from each other. Excluding AILoss leads to the most significant performance drop, which indicates that our modeled aspect information properly guides the generator and constrains the optimization to avoid falling into sub-optimal results.

Table 3. Results of ablation study.

Model	MReD			Historic Place			Plant		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
SALAS-BART$_{\text{large}}$	**39.17**	**13.60**	**23.72**	**34.53**	**11.19**	**19.85**	**28.62**	**6.75**	**16.39**
w/o encoder probe	39.04	12.96	23.58	33.62	11.04	19.48	28.05	6.58	16.05
w/o decoder probe	38.83	11.00	23.08	33.39	10.76	19.27	27.70	6.30	15.83
w/o AILoss	38.59	10.61	22.93	32.40	10.50	19.01	27.18	6.08	15.66

When only keeping the encoder probe and removing other constraints, the performance of the model can still be improved. That is to say, the generator requires more aspect information to improve its performance by infusing aspect information into the representations of input documents. Meanwhile, the small improvement indicates that guiding the generator in an implicit way, giving aspect-specific expressive power to the encoded representations without explicit supervision on the decoder, is not effective enough for a generator to capture and decode the corresponding aspect information in representations. The significant performance drop, caused by removing the decoder probe, also reflects the same conclusion.

4.6 Analysis and Discussion

The above experimental results confirm the effectiveness of our proposed SALAS, we analyze how SALAS is able to achieve the performance in this section.

Table 4. Results of the probed aspects from documents.

Model	Micro			Macro		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
SALAS	**85.64**	**95.58**	**90.34**	**66.48**	**77.53**	**70.88**
TWAG	83.91	69.80	76.21	63.66	54.14	57.42

Effectiveness of the Probe. We evaluate the effectiveness of the aspect probe (encoder probe) that probes aspect information in documents to guide the summarization. As shown in Table 4, compared with TWAG which separates aspect classification from generation. Our joint learned encoder probe not only avoids the cascade error but also achieves better performance on aspect detection. We can observe that our encoder probe outperforms TWAG by 13.46/14.13 macro/micro-F1 score, which demonstrates that SALAS captures the scattered aspect information across documents more accurately than TWAG. Since SALAS recovers missing aspect information and uses it to guide the generator, the increasing quality of aspect information explains how SALAS achieves the current SOTA performance.

Effectiveness of Guidance from Aspects. We demonstrate that the performance improvement depends not only on our probed accurate aspect information but also on its effective guidance. As shown in Fig. **??**, strengthening the guidance of aspect information exerts a significant positive effect on the performance of $\text{BART}_{\text{large}}$. Specifically, the parameters of $\text{BART}_{\text{large}}$ are optimized with the objective function proposed in Eq. (12), where the value of λ_2 iterates from 0 to 1 with a step of 0.1. The increasing λ_2 represents the growing strength of guidance (penalizing the model for missing aspect information). We keep other parameters constant to investigate the relationship between guidance and performance. We observe that (1) the guidance improves the performance of models; (2) the improvement fluctuates when the guidance is weak; (3) After the fluctuation, the stronger guidance leads to a better performance, which has leveled out since exceeding a certain strength.

To sum up, we verify the effectiveness of both our proposed probe and the guidance of aspect information, thus explaining the underlying reason why SALAS achieves the SOTA performance and demonstrating the validity of our idea.

5 Conclusion

Multi-document summarization (MDS) is a long-standing task and is challenging due to the requirement of paraphrasing the key information scattered across multiple documents. In this paper, we introduce our supervised aspect learning abstractive summarization (SALAS) model, which captures aspect information to constrain the optimization and aids representation learning. SALAS adopts a multi-task joint learning method to avoid introducing the cascade error and impeding the interaction between aspect detection and generation. The extracted aspect information guides the generating process, improving the comprehensiveness and faithfulness of the generated summaries. The experimental results on three commonly used summarization datasets not only show that SALAS outperforms the strong baseline models but also validate the effectiveness of the probed aspects which are accurate and well guide the generating process.

References

1. Ahuja, O., Xu, J., Gupta, A., Horecka, K., Durrett, G.: ASPECTNEWS: aspect-oriented summarization of news documents. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6494–6506. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.449
2. Angelidis, S., Amplayo, R.K., Suhara, Y., Wang, X., Lapata, M.: Extractive opinion summarization in quantized transformer spaces. Trans. Assoc. Comput. Linguist. **9**, 277–293 (2021). https://doi.org/10.1162/tacl-a-00366
3. Arora, R., Ravindran, B.: Latent dirichlet allocation and singular value decomposition based multi-document summarization. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 713–718. IEEE, Pisa, Italy (2008). https://doi.org/10.1109/ICDM.2008.55
4. Beltagy, I., Peters, M.E., Cohan, A.: LongFormer: the long-document transformer. ArXiv (2020). https://doi.org/10.48550/ARXIV.2004.05150
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.N.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2018)

7. Erkan, G., Radev, D.R.: LexRank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004). https://doi.org/10.1613/jair.1523
8. Grail, Q., Perez, J., Gaussier, E.: Globalizing BERT-based transformer architectures for long document summarization. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1792–1810. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.eacl-main.154
9. Hayashi, H., Budania, P., Wang, P., Ackerson, C., Neervannan, R., Neubig, G.: WikiAsp: a dataset for multi-domain aspect-based summarization. Trans. Assoc. Comput. Linguist. **9**, 211–225 (2021). https://doi.org/10.1162/tacl-a-00362
10. Hu, X., Wen, L., Xu, Y., Zhang, C., Yu, P.: SelfORE: self-supervised relational feature learning for open relation extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3673–3682. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.299
11. Jin, H., Wang, T., Wan, X.: Multi-granularity interaction network for extractive and abstractive multi-document summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6244–6254. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.556
12. Kiyoumarsi, F.: Evaluation of automatic text summarizations based on human summaries. Proc. Soc. Behav. Sci. **192**, 83–91 (2015). https://doi.org/10.1016/j.sbspro.2015.06.013
13. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880 (2020). https://doi.org/10.18653/v1/2020.acl-main.703
14. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004)
15. Liu, Y., Lapata, M.: Hierarchical transformers for multi-document summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5070–5081. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1500
16. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3728–3738. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1387
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Ma, C., Zhang, W.E., Guo, M., Wang, H., Sheng, Q.Z.: Multi-document summarization via deep learning techniques: a survey. ACM Comput. Surv., 3529754 (2022). https://doi.org/10.1145/3529754
19. Mao, Z., et al.: DYLE: dynamic latent extraction for abstractive long-input summarization. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1687–1698. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.118

20. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004)
21. Over, P., Yen, J.: An introduction to DUC-2004. National Institute of Standards and Technology (2004)
22. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014). https://doi.org/10.3115/v1/D14-1162
23. Radev, D.: A common theory of information fusion from multiple text sources step one: cross-document structure. In: 1st SIGdial Workshop on Discourse and Dialogue, pp. 74–83 (2000). https://doi.org/10.3115/1117736.1117745
24. Shen, C., Cheng, L., Zhou, R., Bing, L., You, Y., Si, L.: MReD: meta-review dataset for structure-controllable text generation. In: Findings of the Association for Computational Linguistics: ACL 2022, pp. 2521–2535. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.findings-acl.198
25. Wang, W., Pan, S.J., Dahlmeier, D., Xiao, X.: Recursive neural conditional random fields for aspect-based sentiment analysis. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 616–626 (2016). https://doi.org/10.18653/v1/D16-1059

26. Xu, J., Durrett, G.: Neural extractive text summarization with syntactic compression. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3290–3301. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1324
27. Yao, L., Haghighi, A., Riedel, S., McCallum, A.: Structured relation discovery using generative models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1456–1466. EMNLP '11, Association for Computational Linguistics, USA (2011)
28. Zhang, Y., et al.: An exploratory study on long dialogue summarization: what works and what's next. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 4426–4433. Association for Computational Linguistics, Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.findings-emnlp.377
29. Zhu, C., Xu, R., Zeng, M., Huang, X.: A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 194–203. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.19
30. Zhu, F., Tu, S., Shi, J., Li, J., Hou, L., Cui, T.: TWAG: a topic-guided wikipedia abstract generator. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4623–4635. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.356