Zero-Shot Aspect-Based Scientific Document Summarization using Self-Supervised Pre-training

Anonymous ACL submission

Abstract

We study the zero-shot setting for the aspect-based scientific document summarization task. Summarizing scientific documents with respect to an aspect can remarkably improve document assistance systems and readers experience. However, existing large-scale datasets contain a limited variety of aspects, causing summarization models to over-fit to a small set of aspects. We establish baseline results in zero-shot performance (over unseen aspects and the presence of domain shift), paraphrasing, leave-one-out, and limited supervised samples experimental setups. We propose a self-supervised pre-training approach to enhance the zero-shot performance. Experimental results on the FacetSum and PubMed aspect-based datasets show promising performance when the model is pre-trained using unlabeled in-domain data.[1]

1 Introduction

Scientific document summarization aims to summarize research papers, and it is usually considered as generating paper abstracts (Cohan et al., 2018). Compared to the news summarization datasets like CNN/Daily Mail (Hermann et al., 2015) and XSUM (Narayan et al., 2018), scientific papers are significantly longer, follow a standard structure, and contain more technical terms and complex concepts (Yu et al., 2020). Recently, there have been remarkable improvements in the area of scientific document summarization due to the availability of large-scale datasets such as arXiv and PubMed (Cohan et al., 2018) and pre-trained sequence to sequence models such as BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020). However, little research has been conducted on aspect-based scientific document summarization.

Aspect-based summarization is the task of summarizing a document with respect to a specific point of interest. Aspect-based scientific document summarization has several advantages for readers to explore retrieved articles quickly and facilitates document assistance systems. It can be particularly helpful to assist readers in critical reviewing of articles (Yuan et al., 2021a). Collecting a large-scale dataset for this task is extremely costly. Meng et al. (2021) introduce FacetSum, an aspect-based document summarization dataset. They employ structured abstracts from the Emerald database[2] to create summaries from four perspectives (*purpose*, *method*, *findings*, and *value*). However, in real applications, readers may be interested in new aspects that go beyond proposed annotations.

Summarization problem heavily relies on sequence-to-sequence models that require large amount of training data. While scientific summa-

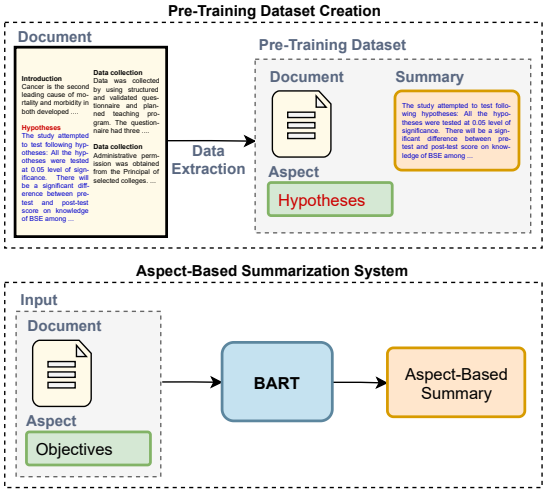

Figure 1: Overview of our approach to create self-supervised pre-training datasets from unlabeled scientific documents. The aspect-based summarization model is pre-trained on unlabeled documents, the section and sub-section headings as aspects, and the following paragraphs corresponding to the aspects as aspect-based summaries.

[1] We will release our dataset and models upon acceptance.

[2] www.emerald.com

rization problem can benefit from large amount of articles with their summaries available (Cohan et al., 2018), the data for aspect-based summarization of scientific literature is scarce. Moreover, most of existing methods for aspect-based summarization rely on pre-defined aspects. Addition of new aspects would require gathering new data and retraining the whole system.

In this work, we are interested in zero-shot aspect-based summarization of scientific literature. Large pre-trained models such as BERT (Devlin et al., 2019) and BART have demonstrated the high potential of knowledge transfer from self-supervised tasks to downstream tasks, up to the emergence of zero-shot capability allowing to solve tasks through "prompting" for very large models (Brown et al., 2020). Continuing the BART pre-training task (i.e., token masking and deletion, text infilling, sentence permutation, and document rotation) with domain-related or target datasets can improve the final performance on low-resource domains. However, this process, specifically using domain-related datasets, is substantially time-consuming (Yu et al., 2021). Also, training a summarization model using a second summarization dataset on the same task (i.e., intermediate training) enhances the performance (Yu et al., 2021). Such approaches only cover limited aspects. We believe that a good aspect-based summarization system should establish semantic similarity between the aspect and the content of the document. The contributions of this work are the following:

- We establish baselines for aspect-based summarization on two different datasets and analyse the zero-shot capabilities of those models on unseen aspects.
- For zero-shot capabilities, we study the effect of domain shift and unseen aspects on aspect-based summarization performance.
- We propose self-supervised pre-training to boost the zero-shot capability of the aspect-based summarization model and demonstrate its effectiveness.
- Finally, we analyse how different models behave as the amount of supervision decreases.

2 Related Work

Abstractive Summarization. Early research on abstractive summarization mainly focused on paraphrasing-based compression methods (Filippova, 2010; Berg-Kirkpatrick et al., 2011). Later motivated by the success of neural attention mechanism in machine translation (Bahdanau et al., 2014), attention-based sequence-to-sequence models have been developed for abstractive summarization (Rush et al., 2015; Nallapati et al., 2016). Adopting the pre-training transformer-based models by self-supervised objectives has led to significant improvements in NLP (Devlin et al., 2019). In particular, BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020) extend such idea to text generation and have the state of the art performance on the abstractive summarization task.

Scientific Document Summarization. Scientific document summarization falls under the problem of long document summarization. Different approaches have been proposed to alleviate models struggle with long inputs, such as applying a hierarchical encoder together with a decoder attending to discourse-level information (Cohan et al., 2018) or summarizing papers sections separately (Gidiotis and Tsoumakas, 2019).

There are several attempts to use Transformer (Vaswani et al., 2017) for long document summarization such as splitting inputs into blocks and applying Transformer layers with shared parameters followed by an extra attention layer to compress sequences into a shorter sequence (Xie et al., 2020). Two-step pipelines (extract relevant information then summarize) is another approach (Yuan et al., 2021b; Gidiotis and Tsoumakas, 2020) to address this problem. BART can also handle long sequences using a hierarchical attention model (Rohde et al., 2021) or simply by extending its positional embedding (Meng et al., 2021). We performed some initial experiments by extending input length processed by BART beyond its default values and found no significant improvement considering extra complexity. Moreover, our initial experiments exposed similar trends across different BART versions. Therefore in follow up experiments, we stick to the standard BART model.

Aspect-based Summarization. Prior to scientific documents, aspect-based summarization has been primary studied on online reviews to summarize opinions (Titov and McDonald, 2008; Lu et al., 2009; Yang et al., 2018; Angelidis and Lapata, 2018), arguments (Wang and Ling, 2016), and news articles (Frermann and Klementiev, 2019; Kr-

PubMed	**# Samples (Aspect, Document)**				
	Train: 139.4K / Validation: 7.9K / Test: 8.1K				
	Average Length (# Words)				
	Documents: 3.5K				
	Summaries:				
	Intro.	Objectives	Methods	Results	Conc.
	53	38	76	94	40
FacetSum	**# Samples (Aspect, Document)**				
	Train: 182.4K/ Validation: 23.7K / Test: 23.7K				
	Average Length (# Words)				
	Documents: 6.6K				
	Summaries:				
		Objectives	Methods	Results	Value
		53	49	66	46

Table 1: Statistics of the PubMed and FacetSum aspect-based scientific summarization datasets.

ishna and Srinivasan, 2018). PMC-SA (Gidiotis and Tsoumakas, 2019) leverages structured scientific abstracts for structured summarization over three sections. In particular, FacetSum (Meng et al., 2021), an aspect-based scientific document summarization, has been collected using the structured outline of the scientific papers crawled from the Emerald database. It covers a wide range of domains but mainly includes marketing, management, education, and economics.

Training separated models per aspects (Hayashi et al., 2020) is not preferable within the zero-shot setting. To integrate the representations of aspect words and input sequences, specific attention mechanism over aspects is used for RNN-based networks (Yang et al., 2018), pointer-generator networks (Krishna and Srinivasan, 2018; Frermann and Klementiev, 2019), and Transformer (Xie et al., 2020). Simple concatenating aspects with documents is a straightforward method result in promising performance using BERT (Xu and Lapata, 2021) and BART (Meng et al., 2021; Tan et al., 2020; Su et al., 2021). In this work, we follow this direction and study to what extent such models are robust to new aspects and domain shift.

Zero-Shot Summarization Hua and Wang (2017) combine in-domain and out-of-domain datasets to improve abstractive summarization on small data. While Magooda and Litman (2020) propose a template-based data synthesis method to incorporate into training that improves the small dataset abstractive summarization. Coavoux et al. (2019) study an entire unsupervised aspect-based abstractive summarization approach but it is difficult to extend this work to predefined aspects. Recently, AdaptSum (Yu et al., 2021) leverages the idea of a second pre-training on BART. They compare intermediate training using a second summarization dataset with continuing the BART pre-training using two pre-training approaches: a time-consuming domain-adaptive pre-training (using a corpus related to the target domain) and task-adaptive pre-training (using the unlabeled target domain). They show intermediate training surpasses continuing the BART pre-training. Similar to our idea of using task-specific self-supervised pre-training, self-supervised generic summaries extracted from the first sentences of Wikipedia documents (Fabbri et al., 2021) and news articles (Zhu et al., 2021) are used to pre-train summarization models for social media, patent document, and news summarization tasks. To the best of our knowledge, our paper is the first study investigating zero-shot aspect-based summarization.

3 Methods

In this section, we first present how we formulate the aspect-based summarization problem relying on BART pre-trained model. Then, we propose a method to use unlabeled data for an additional self-supervised pre-training step to improve the zero-shot performance.

3.1 Aspect-Based Summarization

Given an aspect phrase $A = \{A_1, A_2, ..., A_K\}$ containing K words, and a document $D = \{W_1, W_2, ..., W_N\}$ containing N words, the aspect-based summarization task aims to summarize D into summary $S = \{S_1, S_2, ..., S_M\}$ with respect to aspect A using an autoregressive summarization model $S_{t+1} = Model(S_t, X = \{D, A\})$ for $t = \{0, ..., M-1\}$. We use BART, a pre-trained model combining bidirectional and autoregressive transformers, to encode documents and aspects together and generate aspect-based summaries. To combine aspects and documents as input X, we concatenate A to the beginning of D with the following format:

$$X = <s> \{A_1, ..., A_K\} </s> \{W_1, ..., W_N\}$$

where $<s>$ and $</s>$ are the beginning of sentence, and separation tokens, respectively. Finally, we train the model with cross-entropy loss function similar to a generic summarization task.

3.2 Self-Supervised Training

A model is able to extend its prediction to unseen aspects only if it is able to make a semantic connec-

tion between the aspect and the content of the document. In order to make such connection stronger, the model needs larger and more varied amount of samples than what existing aspect-based datasets make available. In order to extend it, we propose self-supervised pre-training on section and subsection headings from the full articles. We assume headings are phrases conveying the central topic of sections and are good alternatives for aspects.

We propose to extract self-supervised samples from the training set of the PubMed and FacetSum datasets. Figure 1 briefly explains our extraction method. We use the sections and subsections headings as aspect words and phrases. We assign sentences in the corresponding sections or subsections as target summary for each aspect. We truncate the sentences up to a word finishes after the 300th character. Then, we pre-train BART with the extracted dataset using the same cross-entropy loss function we use for training the final summarization task.

We assume training a model to generate relevant sentences conditioned on an aspect (section heading) and a document can improve the model to learn the concept of aspect and conditional text generation and learn representations better for diverse aspects together with documents. In other words, instead of directly training on labelled aspect-based summarization, we train the model indirectly using a self-supervised approach.

4 Datasets

For our experiments, we consider FacetSum (Meng et al., 2021), a faceted (aspect-based) summarization benchmark built on Emerald journal articles. In addition, inspired by FacetSum, we process PubMed (Cohan et al., 2018) and convert it into a large aspect-based scientific document summarization dataset. We scraped the PubMed website to collect the structured abstracts corresponding to the documents in the PubMed summarization dataset. We leverage the structured format of papers abstracts on their web-page to extract five aspects: *introduction*, *objectives*, *methods*, *results*, and *conclusion*. Note, we manually checked the extracted aspects and set rules to converted different spellings, typos, and representations (e.g., *intro→introduction*, *method→methods*) into the five standard aspects. Table 1 shows the datasets statistics. We slightly change the aspects in FacetSum to make it more similar to our dataset and make it possible to study domain shift (*purpose→objectives*, *method→methods*, *findings→results*). This newly created dataset will be released to boost work on aspect-based summarization and improve reproducibility of the results.

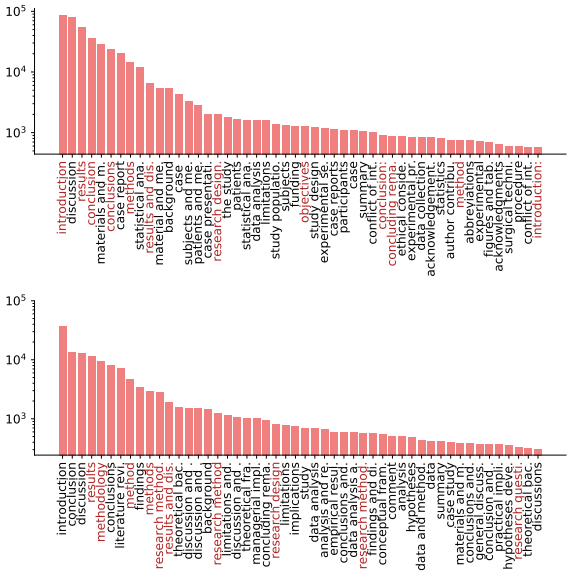

Figure 2: Histogram of 50 most frequent aspects in the self-supervised samples for PubMed* (top) and FacetSum* (bottom). PubMed* has [150069, 1452, 214, 33] unique aspects with frequency of higher than [1, 10, 100, 1000] while FacetSum* has [96525, 841, 120, 21] unique aspects. Aspects removed from the NoOverlap variants of the datasets are highlighted in red.

For self-supervised pre-training step we create two self-supervised datasets: *PubMed** and *FacetSum**, from PubMed and FacetSum aspect based summarization datasets as described in section 3.2. PubMed* and FacetSum* contain 658K and 279K samples and over 150K and 96K unique aspects, respectively. PubMed papers contain more section and subsection headings. Additional dataset PubMed*-NoOverlap and FacetSum*-NoOverlap are the variants of PubMed* and FacetSum* respectively where we exclude headings (aspects) that overlap with the main aspects (shown by red in Figure 2). PubMed*-NoOverlap and FacetSum*-NoOverlap contain 420K and 234K training samples, respectively. Figure 2 shows the distribution of top 50 frequent aspects from PubMed* and FacetSum*.

5 Experiments and Results

In this section, we first explain model hyperparameters. Then, we discuss different experimental setups and analyze results.

We rely on BART base model available through

	Model	R-1	R-2	R-L
PubMed Generic	Discourse (Cohan et al., 2018)	38.93	15.37	35.21
	PEGASUS (Zhang et al., 2020)	39.98	15.15	25.23
	BART	45.04	18.45	40.62
PubMed	Greedy Extractive (Oracle)	56.61	39.23	47.58
	BART	39.03	18.47	34.10
	BART-Independent†	38.91	18.21	33.89
	BART Shuffle Aspects	24.21	6.18	19.86
FacetSum Generic	BART (Meng et al., 2021)	45.49	18.10	42.74
	BART-Facet (Meng et al., 2021)	49.29	19.60	45.76
	BART	49.98	19.89	46.68
FacetSum	Greedy Extractive (Oracle)	51.87	32.09	41.55
	BART (Meng et al., 2021)	23.27	10.31	20.29
	BART-Facet (Meng et al., 2021)	37.97	15.17	32.08
	BART	36.97	15.50	31.48
	BART-Independent†	36.77	15.26	31.23
	BART Shuffle Aspects	28.18	6.94	22.71

Table 2: Baselines and the state of the art performance on PubMed and FacetSum generic and aspect-based summarization evaluation sets. Results for the models with † are averaged over all aspects. Results reported by Meng et al. (2021) are based on a BART model extended to 10,000 tokens

HuggingFace's Transformers library (Wolf et al., 2019). It is then trained for each of the tasks/datasets we tackle. Fine-tuning is done on 1 GPU (NVIDIA V100), with batch size of 64 (8 training samples per GPU and 8 gradient accumulation steps). We train the model for 10 epochs (2 epochs for self-supervised pre-training) with a learning rate of $3e-4$ and 500 warm-up steps and set the maximum input length to 1024, the BART official length.

5.1 Baselines Experiments

System performance is evaluated with the ROUGE metric (Lin and Hovy, 2003). Table 2 reports R-1, R-2 and R-L scores, measuring the N-gram overlap between the reference and generated summaries, for different baseline models evaluated on PubMed and FacetSum datasets. The first part of the table reports the results on the generic summarization task (summarizing documents into full abstracts) for a sanity check and compare the ROUGE scores between *off-the-shelf* BART model, as well as the BART model fine-tuned on PubMed or FacetSum dataset.[3] For aspect-based summarization we consider following baselines:

- *Greedy extractive*: an extractive summarization oracle using the greedy extractive (Nallapati et al., 2017) method. We calculate the ROUGE-N scores (R-1, R-2, and R-L) between every sentence in a document and the reference aspect-based summaries to find top sentences with the highest scores for each document. Next, the best set of top sentences in terms of ROUGE-N scores is selected per document, and then scores are aggregated for all samples. The same score chooses sentences for each ROUGE-N score oracle.

- *BART*: BART model fine-tuned on the aspect-based summarization task containing all the available aspects.

- *BART-Independent*: BART model trained on each aspect independently; we report an average performance across all the aspects. Note that this baseline is not applicable in zero-shot settings.

- *BART-Shuffle*: We evaluate the BART generated aspect-based summaries generated from a wrong aspect (input document is the same but aspects' summaries are replaced randomly, e.g., *objectives→methods*). This baseline serves as a lower-bound of aspect-based summarization performance.

Table 2 shows the baseline results of the generic and aspect-based summarization models. As expected, *greedy extractive* establishes a maximum oracle extractive summarization performance. BART slightly surpasses *BART-Ind*, showing that training all aspects together results in a better performance. Also, independent training is not applicable in the zero-shot setups. *BART-Shuffle* performs significantly worse than the other models. It indicates that the aspects belonging to a specific paper still demand significantly different summaries. Such a model primarily generates generic summaries rather than aspect-related summaries.

Tables 3 and 4 report the performance in terms of different aspects. In both datasets, *objective*-aspect reaches the best ROUGE scores while the performance drops for *results*, *conclusion*, and *value* aspects. Similar phenomenon has been observed by Meng et al. (2021) and can possibly happen due to fact that information needed for summarizing *results*, *conclusion*, and *value* are mostly spread at the end of papers while information about *objectives* is skewed toward the beginning of the papers.

[3] We use the BART model with a length of 1024. We experimented with longer BART models (extending the BART positional embedding to 2,048 and 4,096 tokens) as well as PEGASUS. However, We did not see a significant difference, and therefore we continued all the experiments with the standard BART model.

Model	Introduction	Objectives	Methods	Results	Conclusion
Greedy-Ext.	55.54/38.51/47.09	57.86/37.94/49.65	57.86/37.94/49.65	56.59/40.00/46.09	61.08/44.88/53.81
BART	40.66/22.12/36.18	51.45/31.79/46.09	40.78/19.08/35.84	34.73/12.91/30.69	34.03/14.11/28.17
BART-Ind.	40.76/22.03/36.22	51.11/31.09/45.44	41.01/19.26/35.99	34.16/12.40/30.10	33.95/13.76/28.13
BART-Shuf.	26.14/07.14/21.63	27.94/08.51/22.04	24.07/06.14/19.86	20.16/04.08/17.08	24.67/05.78/19.79

Table 3: Baseline and SOTA performance on the PubMed aspect-based summarization dataset (R-1/R-2/R-L).

Model	Objectives	Methods	Results	Value
Greedy-Ext.	54.94/34.27/44.54	49.27/29.82/39.18	53.25/34.35/42.49	50.18/29.97/40.33
BART (Meng et al., 2021)	46.74/27.09/41.21	23.66/07.92/20.53	16.39/04.63/14.33	06.30/01.62/05.07
BART-Facet (Meng et al., 2021)	48.65/27.72/42.55	33.49/11.01/28.07	34.46/10.49/28.98	35.27/11.44/28.70
BART	48.83/29.10/43.46	32.79/11.71/27.64	32.67/10.21/27.43	33.58/10.98/27.38
BART-Ind.	48.77/28.92/43.31	32.59/11.61/27.39	32.26/09.80/26.96	33.47/10.73/27.26
BART-Shuf.	32.52/09.75/26.34	25.86/05.71/20.96	25.76/05.61/20.83	28.48/06.63/22.79

Table 4: Baseline and SOTA performance on the FacetSum aspect-based summarization dataset (R-1/R-2/R-L).

PubMed					FacetSum				
Pre-Train	**Train**	**R-1**	**R-2**	**R-L**	**Pre-Train**	**Train**	**R-1**	**R-2**	**R-L**
Unlabelled Data									
PubMed*	-	30.76	11.64	26.16	FacetSum*	-	28.18	7.60	23.54
PubMed* (No Overlap)	-	29.70	10.93	25.20	FacetSum* (No Overlap)	-	26.90	6.67	22.45
FacetSum*	-	28.68	9.79	24.30	PubMed*	-	27.24	7.01	22.34
Unlabelled & Out-Of-Domain Labelled Data									
-	FacetSum	28.89	10.20	24.52	-	PubMed	31.03	10.04	25.75
PubMed*	FacetSum	31.31	11.53	26.79	FacetSum*	PubMed	31.67	10.34	26.25
PubMed* (No Overlap)	FacetSum	30.37	10.68	25.69	FacetSum* (No Overlap)	PubMed	31.17	10.10	25.90
FacetSum*	FacetSum	28.92	10.12	24.46	PubMed*	PubMed	30.48	9.48	25.29

Table 5: Performance on the PubMed and FacetSum aspect-based summarization dataset when no labelled data is available or only out-of-domain data is available for intermediate training. PubMed* and FacetSum* are the self-supervised datasets for pre-training.

The performance drop could be also due to the fact that we truncate documents into a maximum sequence length (1024 tokens) required by default BART architecture.

5.2 Zero-Shot Experiments

Zero-shot experiments are reported in Table 5. We define different experimental setups concerning the dataset used for the pre-training and training phase. To be a zero-shot experiment, a model cannot be trained on in-domain labelled dataset. However, it can be pre-trained on the same unlabeled in-domain dataset (PubMed* or FacetSum*) in the self-supervised approach. This is a real-life practical case where there are numerous unlabeled datasets and no labelled samples. As shown in Table 5, the best performance on both datasets belongs to the case that the model is pre-trained on the same but unlabeled dataset, PubMed* or FacetSum*, and fine-tuned the other dataset, PubMed or FacetSum. Also, in-domain pre-training can improve the performance of models that have later an intermediate-training step. This experiment shows that pre-training models using our proposed self-supervised approach by the unlabeled in-domain dataset is a promising approach to improve zero-shot performance. Interestingly, the models pre-trained on the PubMed* dataset performs drastically better on the PubMed dataset than the model, which is only fine-tuned on FacetSum* while this does not hold for the same case on the FacetSum experiment. We hypothesize that it might be due to the significantly larger size of the PubMed* dataset (658K) compared to the FacetSum* dataset (279K). It is also promising that pre-trained models with no overlapping with the target aspect perform quite well. Such cases simulate the entirely new and unseen aspects in real scenarios.

5.3 Leave-One-Out Experiments

This section studies leave-one-out experiments, aiming to investigate performance on unseen aspects within the same domain. We fine-tune BART for aspect-based summarization on all aspects except one that is left out for evaluation. We repeat the experiments for all the aspects available within our dataset. Table 6 reports the results for this ex-

			PubMed			FacetSum		
Pre-Train	**Train**	**Test**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
✗	All - Introduction	Introduction	30.88	11.65	25.66	-	-	-
✓	All - Introduction	Introduction	40.07	21.22	35.5	-	-	-
✓✓	All - Introduction	Introduction	38.76	20.29	33.86	-	-	-
✗	All - Objectives	Objectives	28.97	8.97	22.99	29.08	8.33	23.87
✓	All - Objectives	Objectives	34.28	14.26	28.06	36.28	12.92	29.74
✓✓	All - Objectives	Objectives	30.69	10.60	24.84	29.15	8.28	23.77
✗	All - Methods	Methods	25.68	7.03	21.10	27.32	6.59	22.16
✓	All - Methods	Methods	27.28	7.70	22.23	28.13	6.84	22.79
✓✓	All - Methods	Methods	27.41	7.89	22.8	28.07	6.59	22.63
✗	All - Results	Results	21.28	4.68	17.92	23.82	5.25	19.47
✓	All - Results	Results	22.86	5.05	19.51	23.07	4.80	18.90
✓✓	All - Results	Results	21.12	4.67	17.79	24.22	5.28	19.83
✗	All - Conclusion	Conclusion	27.92	7.36	21.86	-	-	-
✓	All - Conclusion	Conclusion	31.23	9.17	24.73	-	-	-
✓✓	All - Conclusion	Conclusion	30.03	8.13	23.49	-	-	-
✗	All - Value	Value	-	-	-	30.41	7.86	24.22
✓	All - Value	Value	-	-	-	31.45	7.92	25.05
✓✓	All - Value	Value	-	-	-	29.25	7.41	23.52

Table 6: Leave-one-out experiment on the PubMed and FacetSum aspect-based summarization datasets. The models are trained on all aspects except the one which the model is tested on. ✗: no pre-training except the BART official pre-training. ✓: model is pre-trained on PubMed* or FacetSum* (in-domain). ✓✓: model is pre-trained on PubMed* (No Overlap) or FacetSum* (No Overlap) (in-domain).

		PubMed			FacetSum		
Pre-Train	**Paraphrased Aspect**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
✗	Introduction	40.66	22.12	36.18	-	-	-
✗	Introduction -> Background	27.98	9.34	23.62	-	-	-
✗	Introduction -> Context	30.37	11.92	25.95	-	-	-
✓	Introduction -> Background	41.47	22.48	36.79	-	-	-
✓	Introduction -> Context	40.28	21.58	35.64	-	-	-
✗	Objectives	51.45	31.79	46.09	48.83	29.10	43.46
✗	Objectives -> Objective	51.37	31.66	46.03	48.91	29.17	43.52
✗	Objectives -> Purpose	36.03	15.93	29.84	46.70	26.11	41.11
✗	Objectives -> Aims	28.89	9.29	23.02	30.95	9.64	25.34
✓	Objectives -> Objective	51.10	31.39	45.60	48.51	28.81	43.14
✓	Objectives -> Purpose	49.77	29.92	44.09	48.28	28.46	42.88
✓	Objectives -> Aims	42.67	22.99	36.72	45.19	24.82	39.55
✗	Methods	40.78	19.08	35.84	32.79	11.71	27.64
✗	Methods -> Method	40.67	18.75	35.753	32.94	11.82	27.73
✗	Methods -> Materials and Methods	40.84	19.16	35.82	32.98	11.75	27.82
✗	Methods -> Research Design	34.82	14.23	29.74	32.68	11.34	27.41
✗	Methods -> Methodology	40.88	19.13	35.90	32.92	11.82	27.81
✓	Methods -> Method	41.13	19.24	36.07	32.85	11.88	27.69
✓	Methods -> Materials and Methods	40.58	19.05	35.58	32.77	11.80	27.69
✓	Methods -> Research Design	38.22	17.18	33.12	32.84	11.81	27.62
✓	Methods -> Methodology	40.82	19.24	35.75	32.77	11.82	27.62
✗	Results	34.73	12.91	30.69	32.67	10.21	27.43
✗	Results -> Result	34.42	12.73	30.30	32.46	10.05	27.21
✗	Results -> Discussion	23.57	7.09	20.09	26.12	5.90	21.25
✗	Results -> Finding	24.85	6.01	21.37	26.63	6.40	21.81
✓	Results -> Result	34.12	12.53	30.00	32.46	9.98	27.22
✓	Results -> Discussion	19.80	4.18	16.65	29.06	7.82	23.93
✓	Results -> Finding	29.11	9.24	25.29	32.46	10.01	27.20
✗	Conclusion	34.03	14.11	28.17	-	-	-
✗	Conclusion -> Conclusions	33.97	14.13	28.16	-	-	-
✓	Conclusion -> Conclusions	33.94	13.92	28.04	-	-	-
✗	Value -> Value	-	-	-	33.58	10.98	27.38
✗	Value -> Values	-	-	-	32.24	10.59	26.98
✓	Value -> Values	-	-	-	33.46	10.99	27.35

Table 7: Paraphrasing experiments performance on the PubMed and FacetSum aspect-based summarization datasets. In each section, we evaluate the model trained on all original aspects on a new paraphrased aspect, e.g., *introduction→background* reports the case when *introduction* summaries are assigned to *background*.

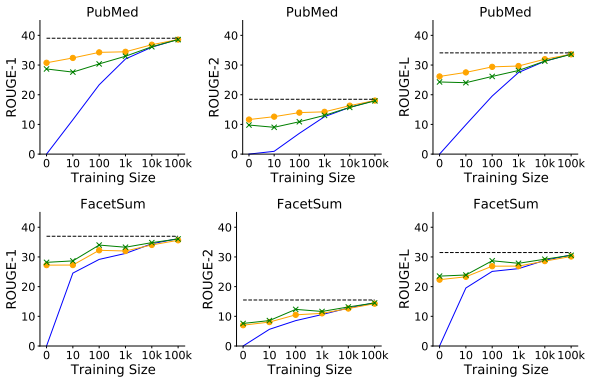

Figure 3: Aspect-based summarization performance with limited supervised examples. Pre-training with in-domain and out-of-domain datasets significantly improves the low-resource training sample performance. Top: evaluation done on PubMed dataset, Bottom: evaluation is done on FacetSum dataset. (—— *BART*, –●– *BART + pre-trained on PubMed**, –×– *BART + pre-trained on FacetSum**, - - - *BART fine-tuned on all samples*)

periment for both PubMed and FacetSum datasets. We compare baseline model (✗) and models enriched with self-supervised pre-training step as described in the section 3.2. The self-supervision pre-training can be done either on all the section headings (✓) or only on those non-overlapping with aspects of interest (✓✓). First, we note that zero-shot performance without self-supervised pre-training performs significantly worse compared to fully supervised models although it is still above random lower bound BART-Shuffle model (cf. tables 3 and 4). The pre-training step allows to significantly improve this performance for most of the aspects. As shown, non-overlapping pre-training (✓✓) can also increase the performance in most of the cases except *results* and *value*. *introduction* and *objective* experience the most improvement. As discussed previously (section 5.1) this could be due to the fact that information required to summarize these aspects are skewed toward the beginning of papers (Meng et al., 2021), and therefore is always within the input range of BART.

5.4 Paraphrasing Experiments

In this section, we study another zero-shot experiment in which the aspect word or phrase is paraphrased for evaluation. This experiment aims to understand to what extent a model can exploit semantic meaning of aspects to generate good summaries. Table 7 reports results for this experiment comparing models with and without pre-training. As in the previous experiment, we see that the baseline model (without pre-training) may suffer from a significant drop when replacing the original aspect with its alternative. However, it still performs better than the random lower bound model (cf. tables 3 and 4). Again, the pre-training step makes the model more robust to aspects paraphrasing. This is probably due to the fact that the model has been exposed to a much richer and more scarce set of aspects during pre-training, and therefore learned to exploit better aspect phrases.

5.5 Low Resource Experiments

Our final experiment aims at evaluating the summarization performance with limited supervised examples. For this, we train BART on the first 10, 100, 1K, 10K, and 100K training samples from each dataset. We repeat the experiments with the BART models pre-trained on the PubMed* and FacetSum* self-supervised datasets. Figure 3 plots the learning curves behaviour of different models as the amount of supervision grows. We see that models with self-supervised pre-training consistently surpass the baseline model. This superiority is much more significant in the few-shot cases, but the differences fade as more training samples is available. As expected, the models pre-trained on in-domain datasets perform better than the out-domain pre-trained models.

6 Conclusion

In this paper, we studied the problem of zero-shot aspect-based summarization of scientific documents. We established various experimental setups to investigate the effect of additional pre-training and intermediate training on the zero-shot performance with respect to domain-shift and unseen aspects. We proposed a self-supervised approach to pre-train the model using unlabeled target datasets. Results indicate that additional pre-training on the target dataset followed by intermediate training results in the best zero-shot performance.

We established leave-one-out and paraphrasing experimental setups to simulate the practical case of facing unseen aspects and showed the promising effect of additional self-supervised pre-training. Our proposed pre-training step improves the performance in the few-shot settings.

Investigating the effect of pre-training in terms of semantics evaluation scores can be done in the future.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, 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. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online. Association for Computational Linguistics.

Katja Filippova. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 322–330, Beijing, China. Coling 2010 Organizing Committee.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

Alexios Gidiotis and Grigorios Tsoumakas. 2019. Structured summarization of academic publications. *arXiv preprint arXiv:1905.07695*.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2020. WikiAsp: A dataset for multi-domain aspect-based summarization. *arXiv preprint arXiv:2011.07832*.

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*, 28:1693–1701.

Xinyu Hua and Lu Wang. 2017. A pilot study of domain adaptation effect for neural abstractive summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 100–106, Copenhagen, Denmark. Association for Computational Linguistics.

Kundan Krishna and Balaji Vasan Srinivasan. 2018. Generating topic-oriented summaries using neural attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1697–1705, New Orleans, Louisiana. Association for Computational Linguistics.

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

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence

statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140.

Ahmed Magooda and Diane Litman. 2020. Abstractive summarization for low resource data using domain transfer and data synthesis. In *The Thirty-Third International Flairs Conference*.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. *arXiv preprint arXiv:2106.00130*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gůlçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Tobias Rohde, Xiaoxia Wu, and Yinhan Liu. 2021. Hierarchical learning for generation with long source sequences. *arXiv preprint arXiv:2104.07545*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Dan Su, Tiezheng Yu, and Pascale Fung. 2021. Improve query focused abstractive summarization by incorporating answer relevance. *arXiv preprint arXiv:2105.12969*.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309, Online. Association for Computational Linguistics.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yujia Xie, Tianyi Zhou, Yi Mao, and Weizhu Chen. 2020. Conditional self-attention for query-based summarization. *arXiv preprint arXiv:2002.07338*.

Yumo Xu and Mirella Lapata. 2021. Generating query focused summaries from query-free resources. 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)*, pages 6096–6109, Online. Association for Computational Linguistics.

Min Yang, Qiang Qu, Ying Shen, Qiao Liu, Wei Zhao, and Jia Zhu. 2018. Aspect and sentiment aware abstractive review summarization. In *Proceedings of the 27th international conference on computational linguistics*, pages 1110–1120.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5892–5904, Online. Association for Computational Linguistics.

Tiezheng Yu, Dan Su, Wenliang Dai, and Pascale Fung. 2020. Dimsum@ laysumm 20: Bart-based approach for scientific document summarization. *arXiv preprint arXiv:2010.09252*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021a. Can we automate scientific reviewing? *CoRR*, abs/2102.00176.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021b. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*.

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.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1462–1471.