Efficient Few-Shot Fine-Tuning for Opinion Summarization

Arthur Bražinskas[1] **Ramesh Nallapati**[2] **Mohit Bansal**[2,3] **Markus Dreyer**[2]
[1]ILCC University of Edinburgh [2]Amazon
[3]UNC Chapel Hill
abrazinskas@ed.ac.uk, {rnallapa, mobansal, mddreyer}@amazon.com

Abstract

Abstractive summarization models are typically pre-trained on large amounts of generic texts, then fine-tuned on tens or hundreds of thousands of annotated samples. However, in opinion summarization, large annotated datasets of reviews paired with reference summaries are not available and would be expensive to create. This calls for fine-tuning methods robust to overfitting on small datasets. In addition, generically pre-trained models are often not accustomed to the specifics of customer reviews and, after fine-tuning, yield summaries with disfluencies and semantic mistakes. To address these problems, we utilize an efficient few-shot method based on adapters which, as we show, can easily store in-domain knowledge. Instead of fine-tuning the entire model, we add adapters and pre-train them in a task-specific way on a large corpus of unannotated customer reviews, using held-out reviews as pseudo summaries. Then, fine-tune the adapters on the small available human-annotated dataset. We show that this self-supervised adapter pre-training improves summary quality over standard fine-tuning by 2.0 and 1.3 ROUGE-L points on the Amazon and Yelp datasets, respectively. Finally, for summary personalization, we condition on aspect keyword queries, automatically created from generic datasets. In the same vein, we pre-train the adapters in a query-based manner on customer reviews and then fine-tune them on annotated datasets. This results in better-organized summary content reflected in improved coherence and fewer redundancies.

1 Introduction

Online reviews play an important role in purchasing decisions we make. They inform us about customer experiences – what aspects users like and dislike, and ultimately, whether a product or service is worth buying. Although significant progress has been made in supervised summarization in non-subjective single-document context, such as news articles (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Bražinskas et al., 2021b), modern deep learning methods rely on large amounts of annotated data, but these are not readily available in the review or opinion summarization domain and are expensive to produce. A key obstacle making such annotation expensive is that annotators need to consider multiple texts when writing a summary, which can be tens or even hundreds for realistic settings. Hence, most available datasets have human-written summaries for less than 100 products.

The lack of sufficiently large annotated datasets led to a variety of unsupervised abstractive models (e.g., COPYCAT (Bražinskas et al., 2020b), MEANSUM (Chu and Liu, 2019); DENOISESUM (Amplayo and Lapata, 2020)) that are trained on large collections of unannotated customer reviews. However, as the models are never exposed to actual summaries, they cannot capture their expected characteristics. This results in generated summaries mimicking the informal style of customer reviews and containing hallucinations and unimportant details. These limitations were addressed by few-shot methods learning from a handful of human-written summaries. (Bražinskas et al., 2020a; Oved and Levy, 2021). The first proposed few-shot learning model, FEWSUM (Bražinskas et al., 2020a), relies on various static features capturing differences between a customer review and a summary. While that model is more robust to overfitting, such features require manual, domain-specific engineering and can be sub-optimal for capturing correspondences between texts on the semantic level. We propose a simpler approach – ADASUM – which is based on adapters (Houlsby et al., 2019; Bapna et al., 2019). As we explain next, the adapters are pre-trained on customer reviews in a task-specific manner, and subsequently fine-tuned on gold summaries.

Adapters. We utilize a pre-trained model with powerful language understanding and generation

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1509 - 1523
July 10-15, 2022 ©2022 Association for Computational Linguistics

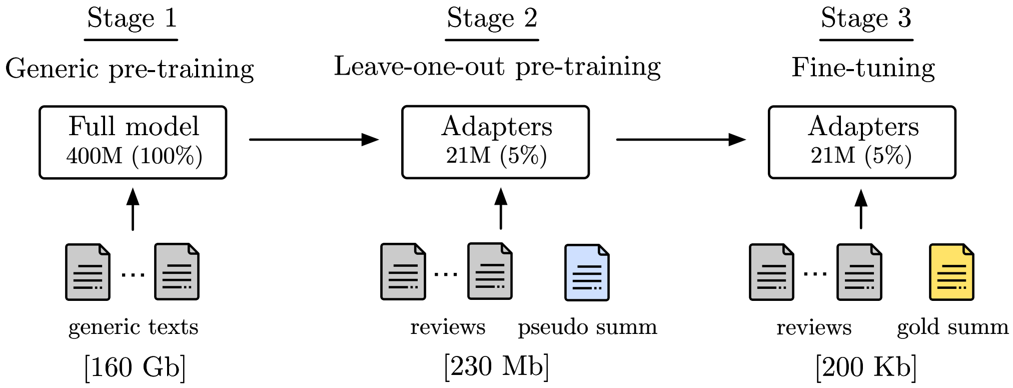

Figure 1: Illustration of the proposed approach. In Stage 1, all parameters of a large language model are pre-trained on generic texts (we use BART). In Stage 2, we pre-train adapters (5% of the full model's parameters) on customer reviews using held-out reviews as summaries. In Stage 3, we fine-tune the adapters on a handful of reviews-summary pairs.

abilities in combination with a parameter-efficient fine-tuning method – adapters. As was shown in recent studies, this method is also robust to overfitting in low-resource settings (He et al., 2021). In this way, a large pre-trained model, BART (Lewis et al., 2020), in our case, remains frozen, and only small modules (0.6% - 5% of the model parameters) are optimized. This effectively retains acquired knowledge in the pre-trained language model (PLM) without specialized training objectives as in RECADAM (Chen et al., 2020). However, available annotated data is not sufficient for learning in-domain specifics and results in summaries with subtle semantic mistakes. As explained next, we reduce these semantic mistakes by pre-training adapters on customer reviews.

Self-supervised Pre-training. Language models are pre-trained with generic objectives (e.g., single-document denoising) and rarely on in-domain data, such as customer reviews. Consequently, this makes them less attuned to in-domain specifics as these are hard to learn from a handful of summaries. This often results in subtle semantic mistakes. For instance, in Table 1, PASS (Oved and Levy, 2021) incorrectly concludes that *thin material* implies *poor quality*. To address this issue, we learn in-domain specifics from customer reviews. Concretely, we employ a self-supervised pre-training method: For any given product without a human-written summary, we predict one of the given reviews by conditioning on N other reviews with the highest lexical overlap in the *leave-one-out* fashion (Bražinskas et al., 2020b). As the standard training of PLMs is storage and memory inefficient (Mahabadi et al., 2021), we pre-train adapters only; see Stage 2 in Fig. 1. Afterwards, we fine-tune them on a small number of annotated reviews-summary pairs (< 100 pairs), see Stage 3 in Fig. 1. All in all, our method combines the general text generation and understanding abilities of the PLM with in-domain knowledge directly related to the end task.

Content Planning. Well-organized content in summaries is easier to follow and thus improves user experience. However, the lack of annotated data makes it challenging to learn a desired content structure. For example, in Table 1, FEWSUM's summary does not end after a concluding phrase '*Other than that, it's a great top.*' While the state-of-the-art model (PASS) addresses this issue by ranking multiple generated candidates with a specialized coherence model, we propose a simpler solution – ADAQSUM – that capitalizes on *text planning* (Hua and Wang, 2019; Moryossef et al., 2019). Specifically, we allow the model to *plan ahead* by providing an intermediate summary representation in the form of a query consisting of aspect keywords. As we show, this results in more coherent text patterns with fewer redundancies. Moreover, it can be useful for personalized summaries, better reflecting user interests.

Result Highlights. We evaluate the proposed models in terms of automatic metrics and human efforts. We find that pre-training and fine-tuning of adapters leads to more than 2.0 and 1.3 ROUGE-L points improvement over fine-tuning the entire model on Amazon and Yelp datasets, respectively. We also find that our pre-trained and fine-tuned query-based model improves ROUGE-L scores by more than 2.7 and 0.9 ROUGE-L points over PASS, on Amazon and Yelp datasets, respectively, and is more preferred by humans. We further demonstrate that the query-based model (ADAQSUM) substantially improves coherence and reduces redundancies in generated summaries.

FEWSUM	This tank top is well made, fits well, and is comfortable to wear. The only thing is that it runs a little small, so order a size up from what you normally wear. Other than that, it's a great top. It's well made and it looks like it will last a long time. Love it!
PASS	This is a basic tank. The photo shows it going well past the models hips. However, the material used to make it this long is thin and therefore not good quality. It is also thinner than other tanks on the market but is still comfortable to wear.
ADAQSUM	This is a basic tank top that fits well and is comfortable to wear. The color is great and the length is long enough to wear with leggings. The quality of the product is good.
REVIEWS	... This is a basic tank ... ‖ ... this tank fits like a normal tank top, not any longer ... I could wear it with leggings ... ‖ ... It is THIN and runs SMALL ... It fits tight and is NOT long like in the picture ... ‖ The tank fit very well and was comfortbale to wear. I've bought much higher quality tanks ... ‖ ... it is listed as a 'long' tank top and the photo even shows it going well past the models hips, however I'm short and the tank top is just a normal length. ‖ ... They were a lot thinner than I like ... ‖ Every women should own one in every color. Just feels quality I don't know how else to explain it ... ‖ ... They are long enough that the color peeks out from under my tops. Looks cute.

Table 1: Generated summaries for an Amazon product by baseline models (FEWSUM and PASS) and our approach (ADAQSUM). Colored words indicate aspect keywords that were part of the query. The special marker '‖' separates truncated reviews.

In summary, our contributions are as follows:

- We propose a self-supervised pre-training method to learn in-domain knowledge by adapters that alleviate catastrophic forgetting;
- We propose, to the best of our knowledge, the first aspect-based abstractive opinion summarizer learned from a few annotated samples;
- We substantially increase summary coherence using the query-based approach;
- We show that self-supervised pre-training significantly improves performance on the query-based task;
- We demonstrate state-of-the-art results on two primary benchmarks in automatic and human evaluation.[1]

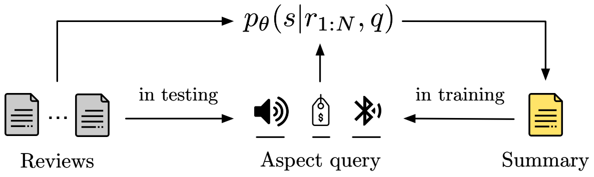

Figure 2: Illustration of the query-based summarizer that inputs reviews and a text query consisting of aspects, such as 'volume,' 'price,' and 'bluetooth.' The query is automatically created from gold summaries in training and reviews in test time.

2 Approach

2.1 Opinion Summarization Tasks

In this work, we consider two tasks of customer review summarization. The first one is *generic summarization* (Chu and Liu, 2019; Bražinskas et al., 2020a), where the aim is to produce a summary that covers overall opinions in input reviews. Formally, given N input reviews $r_{1:N}$, the task is to predict word-by-word the summary s:

$$\mathcal{L}(s, r_{1:N}; \theta) = \sum_{t=1}^{T} \log p_\theta(s^t | s^{1:t-1}, r_{1:N}).$$

In the second task, *query-based summarization*, we assume that the user provides a query q consisting of aspect keywords, such as 'bluetooth,' 'resolution,' and 'battery life.' In turn, a summarizer should generate a summary reflecting customer opinions in $r_{1:N}$ about these aspects. Formally, given a pair of input reviews and query ($r_{1:N}$, q), the task is to predict word-by-word the summary s:

$$\mathcal{L}(s, r_{1:N}, q; \theta) = \sum_{t=1}^{T} \log p_\theta(s^t | s^{1:t-1}, r_{1:N}, q).$$

Unfortunately, abstractive opinion datasets with annotated aspect queries are unavailable in the domain. To mitigate this problem, we follow Ni et al. (2019) and create queries by extracting fine-grained aspect keywords from available generic summaries. Specifically, we utilize the model proposed by Zhang et al. (2014) to build a fine-grained aspect lexicon from review datasets. Further, we use simple rules to determine which aspects appear in summaries; see an annotated summary in Table 2. At test time, we follow the intuition that a summary should reflect common opinions and create a query from K most frequent aspect keywords in input reviews. The workflow is illustrated in Fig. 2.

[1] Our code and associated artifacts will be publicly available at https://github.com/amazon-research/adasum.

2.2 Model

Our model is based on the Transformer (Vaswani et al., 2017) encoder-decoder architecture initialized with BART (Lewis et al., 2020). We adopt the same encoder as in (Raffel et al., 2020; Oved and Levy, 2021) where reviews are concatenated before encoding.[2] This allows us to capture product-level features and leverage commonalities across reviews during encoding. For query-based summarization, we concatenate a query and reviews while indicating boundaries with special markers. In this way, the encoder can contextualize aspect keywords and focus on salient review fragments reflecting these aspects.

2.3 Adapters

In training, a large pre-trained model remains frozen and only tiny neural networks called *adapters* (Houlsby et al., 2019) are optimized. These modules are injected into the transformer layers (both encoder and decoder). Formally, given the input hidden vector h, the output vector $\hat{h}$ is calculated as shown below:

$$\hat{h} = f_2(\tanh f_1(h)) + h.$$

The functions $f_1(\cdot)$ and $f_2(\cdot)$ are the down- and up- projection layers. At each transformer layer, two adapters are inserted right after the self-attention and the feed-forward layers, respectively. These modules consist of substantially fewer parameters than the language model, usually around 3% - 5%. Recent studies have shown that adapters are less prone to overfitting (He et al., 2021) and are more memory-efficient in training (Mahabadi et al., 2021). Finally, as the pre-trained model remains frozen, it retains all the prior knowledge for text understanding and generation. This effectively alleviates catastrophic forgetting (Goodfellow et al., 2013; Kemker et al., 2017) without modifying the training objective as in RecAdam (Chen et al., 2020; Yu et al., 2021). We refer to our approaches as AdaSum and AdaQSum for generic and query-based summarization, respectively.

2.4 Self-supervised Pre-training

Language models, initially pre-trained on generic text corpora, are often not accustomed to in-domain specifics. Unsurprisingly perhaps, a wide range of product-related specifics cannot be learned from a handful of annotated summaries during fine-tuning. Consequently, this can result in subtle semantic mistakes in generated summaries. We will discuss this problem and provide examples in Sec. 5.4. Furthermore, query-based summarization is even more challenging for learning than generic summarization. To be useful in practice, the summarizer should rely on the provided query after fine-tuning. However, a handful of annotated samples might be insufficient to learn this dynamic. We will analyze this problem in Sec. 5.1. To alleviate these two problems, we leverage unannotated customer reviews to construct synthetic datasets for pre-training.

The *cover* offers durable *protection* for the MacBook, the retractable *tilt* stands offer *protection* for the *wrists*.
The *keyboard cover* can take some effort to *fit* properly, and *adjustment* to its feel may take time.
However, free and fast *shipping* make up for this one potential issue.

Table 2: Automatically annotated Amazon summary with fine-grained aspect keywords (*underlined italic*).

Synthetic In-Domain Pre-Training Dataset. From a group of product reviews, we randomly sample one review as a *pseudo summary* s and select N reviews as input ($r_{1:N}$).[3] We select N input reviews covering the content of the summary s – that have the highest ROUGE-1 F scores. Following the naming convention (Bražinskas et al., 2020b), we refer to this as *leave-one-out pre-training* (L1O). To closely resemble query-based summarization, we create aspect queries from pseudo summaries. Specifically, we leverage the aspect lexicon by matching summary keywords; in the same way as was explained in Sec. 2.1. In practice, we expect queries to have at least one aspect keyword. Therefore, we remove all pre-training pairs where the pseudo summary has no aspect keywords.

3 Experimental Setup

3.1 Data

To create synthetic datasets, we used customer reviews from Amazon (He and McAuley, 2016) and

[2] We experimented with the independent review encoding as in Bražinskas et al. (2020b). However, the results were slightly worse.

[3] We also experimented with selecting pseudo summaries without personal pronouns – written in the formal style (Bražinskas et al., 2020a). However, we did not observe significant improvements.

	Amazon		Yelp	
Split	Gold	Synthetic	Gold	Synthetic
Train	84	70,144 / 59,963	90	70,144 / 68,499
Valid	36	7,900 / 6,810	90	7,900 / 7,724
Test	60	-	120	-

Table 3: Source-target pair numbers for Amazon and Yelp, both gold and synthetic. Each pair has 8 source reviews. Generic and query-based pair statistics are separated by '/'.

Yelp.[4] Following Bražinskas et al. (2020a), we selected 4 categories: *Electronics*; *Clothing, Shoes and Jewelry*; *Home and Kitchen*; *Health and Personal Care*. We pre-processed the datasets by removing all reviews that are shorter than 20 words and longer than 120 words and evened the number of pairs in both datasets. Further, we used Amazon and Yelp gold summaries from (Bražinskas et al., 2020a) where each product/business has 3 references and is paired with 8 reviews. Gold and synthetic dataset statistics[5] are presented in Table 3.

3.2 Baselines

LEXRANK (Erkan and Radev, 2004) is an unsupervised extractive graph-based model that selects sentences based on graph centrality. Sentences represent nodes in a graph whose edges are weighted with tf-idf.

MEANSUM (Chu and Liu, 2019) is an unsupervised abstractive summarization model which treats a summary as a structured latent state of an auto-encoder trained to reconstruct reviews of a product.

COPYCAT (Bražinskas et al., 2020b) is the state-of-the-art unsupervised abstractive summarizer with hierarchical continuous latent representations to model products and individual reviews.

FEWSUM (Bražinskas et al., 2020a) is a few-shot framework where lexical features are used to differentiate between customer reviews and summaries. In the fine-tuning phase, features leading to generation of summaries are searched.

PASS (Oved and Levy, 2021) is based on a pre-trained T5 model (Raffel et al., 2020) that is further fine-tuned on gold summaries. At inference, the model's input is perturbed to generate multiple candidates. These candidates are further ranked by a separate model based on coherence and fluency to select the best one.

We fine-tuned the full BART model (FULL) for a fair comparison, with and without the leave-one-out pre-training. We also employed a number of simple summarization baselines. First, the CLUSTROID review was computed for each group of reviews as follows. We took each review from a group and computed ROUGE-L with respect to all other reviews. The review with the highest ROUGE score was selected as the clustroid review. Second, we sampled a RANDOM review from each group to be used as the summary. And lastly, we constructed the summary by selecting the *leading sentences* (LEAD) from each review of a group.

3.3 Experimental Details

We used a standard Transformer encoder-decoder (Vaswani et al., 2017), pre-initialized with BART large (Lewis et al., 2020), consisting of 400M parameters. We used two adapter sizes – 0.6% and 5% of the full model's parameters. All input reviews were concatenated, following Raffel et al. (2020); Oved and Levy (2021). For parameter optimization, we used ADAM (Kingma and Ba, 2014), and summary generation was performed via the beam search of size 5 and with 3-gram blocking (Paulus et al., 2017). We used ROUGE-L as the stopping criterion on the end task, and perplexity (PPL) for pre-training. The learning rate for most experiments was set to 5e-5. Aspect lexicons for query-based summarization contained 2809 and 4013 fine-grained aspects for Amazon and Yelp, respectively. In pre-training and fine-tuning, we shuffled aspects to break temporal dependencies. For fine-tuning on Yelp, we also found it useful to exclude summary aspect keywords that do not appear in input reviews. This approximately matched the number of aspect keywords for Amazon and Yelp. At test time, we selected up to 6 and 5 most frequent aspects for Amazon and Yelp, respectively. All computations were performed on an 8-GPU p3.8-xlarge Amazon instance.

4 Results

4.1 Automatic Evaluation

Table 4 shows results on the Amazon and Yelp test sets for generic summarization. It shows ROUGE F1 scores (Lin, 2004) as a standard measure of informativeness[6] and perplexity (PPL) as a measure of confusion.

[4] https://www.yelp.com/dataset

[5] For the query-based setup, we removed all instances where targets had no aspects.

[6] For consistency with previous works, we used the same Python package (https://github.com/google-research/google-research/tree/master/rouge)

		Amazon				**Yelp**			
	Params↓	PPL↓	R1↑	R2↑	RL↑	PPL↓	R1↑	R2↑	RL↑
Clustroid	-	-	27.16	3.61	16.77	-	28.90	4.90	18.00
Lead	-	-	27.00	4.92	14.95	-	26.20	4.57	14.32
Random	-	-	25.00	3.82	15.72	-	21.48	2.59	13.87
Unsupervised									
LexRank (Erkan and Radev, 2004)	-	-	27.72	5.06	17.04	-	26.96	4.93	16.13
MeanSum (Chu and Liu, 2019)	25M	-	26.63	4.89	17.11	-	27.50	3.54	16.09
Copycat (Bražinskas et al., 2020b)	25M	-	27.85	4.77	18.86	-	28.12	5.89	18.32
Few-shot									
FewSum (Bražinskas et al., 2020a)	25M	-	33.56	7.16	21.49	-	37.29	9.92	22.76
PASS (Oved and Levy, 2021)	440M	-	37.43	8.02	23.34	-	36.91	8.12	23.09
Full (100%)	400M	17.87	37.22	9.17	23.51	12.87	37.40	10.27	23.76
Full (100%) + L1O	400M	16.90	37.67	10.28	24.32	12.40	36.79	11.07	25.03
AdaSum (0.6%)	2.6M	13.45	38.49	9.84	24.37	11.94	37.55	10.11	24.08
AdaSum (0.6%) + L1O	2.6M	12.06	38.94	10.63	24.95	11.23	37.78	11.31	24.04
AdaSum (5%)	21.3M	16.30	38.15	9.18	23.17	12.50	38.12	10.89	24.11
AdaSum (5%) + L1O	21.3M	**12.03**	**39.78**	**10.80**	**25.55**	**11.11**	**38.82**	**11.75**	**25.14**

Table 4: Test set ROUGE F1 scores on gold Amazon and Yelp datasets for generic review summarization. L1O stands for leave-one-out pre-training. We also provide the total number of trainable parameters.

First of all, the results indicate the superiority of adapters over full fine-tuning and state-of-the-art few-shot models on both datasets. As was observed in He et al. (2021), adapters are less prone to overfitting, which is especially beneficial in few-shot settings. Second, we observe a significant improvement in ROUGE scores when pre-trained models are further trained using L1O. This signifies the importance of learning in-domain specifics before fine-tuning. We also observe that adapters are more effective on the Amazon dataset, which is more challenging as indicated by higher perplexity (PPL).[7] We hypothesize that the pre-trained language model (BART) is more accustomed to restaurant- than product-related texts. Moreover, larger adapters (5%) tend to overfit on the small number of annotated instances, and L1O pre-training helps substantially, as indicated both by ROUGE scores and PPL. We provide example generated summaries in the Appendix.

4.2 Human Evaluation

Coherence Improvement. As was observed in (Oved and Levy, 2021), opinion summarizers sometimes generate incoherent summaries. We hypothesized that a query should allow the model to plan ahead of time and thus generate more coherent and less redundant texts. To test the hypothesis, we compared 5% adapter-based models with and without the query; both were pre-trained via L1O. We performed human evaluation in terms of *coherence* and *non-redundancy* via Best-Worst Scaling (BWS) (Louviere and Woodworth, 1991; Louviere et al., 2015). BWS has been shown to produce more reliable results than ranking scales (Kiritchenko and Mohammad, 2016).

For each Amazon test set entry and criterion, we asked three independent workers on Amazon Mechanical Turk (AMT) to select the best and worst summary. For each criterion, a system's score is computed as the percentage of times it was selected as best, minus the percentage of times it was selected as worst (Orme, 2009). The scores range from -100 (unanimously worst) to +100 (unanimously best). For more details, please refer to Appendix 10.2.

First, the results indicate that the summaries generated by AdaQSum are substantially more preferred to AdaSum in terms of coherence and non-redundancy. Namely, +13.73 vs -30.91 and -1.96 vs -25.93 for coherence and non-redundancy, respectively. We also computed the percentage of unique n-grams in each generated summary for both datasets, as shown in Table 5. The results support that query-based summaries are less redundant. However, similar to findings in Oved and Levy (2021), we observe that more coherent summaries tend to get lower ROUGE scores. Nevertheless, our model outperforms PASS by a margin on both datasets – by 2.72 and 0.9 ROUGE-L points on Amazon and Yelp, respectively.

[7] Training sets are of similar sizes, i.e., 84 and 90 summaries on Amazon and Yelp, respectively

		R1	R2	RL	unique 1-gram (%)	unique 2-gram (%)
Amazon	AdaSum (5%) + L1O	**39.78**	**10.80**	25.55	67.72	80.83
	AdaQSum (5%) + L1O	38.53	10.52	**26.06**	**69.38**	**82.57**
Yelp	AdaSum (5%) + L1O	**38.82**	**11.75**	**25.14**	62.26	76.55
	AdaQSum (5%) + L1O	36.79	10.06	23.99	**65.74**	**79.88**

Table 5: Comparison of the query-based and generic summarizers on test sets. Unique n-grams were computed in generated summaries.

	Fluency	Coher.	Non-Red.
PASS	-21.74	**+33.33**	0.00
LexRank	-45.95	-52.38	-58.97
AdaQSum (5%) + L1O	**+26.67**	+25.00	**+26.67**
Gold	+46.67	+27.78	+55.56

Table 6: Human evaluation results in terms of the Best-Worst scaling on the Amazon test set.

	Full↑	Partial↑	No↓
FewSum	47.56	24.39	28.05
PASS	60.70	31.84	7.46
AdaSum (5%) + L1O	**78.97**	**15.48**	**5.56**
AdaQSum (5%) + L1O	72.69	20.37	6.94

Table 7: Input fidelity on the Amazon test set, normalized by sentences.

Comparison to Baselines. To understand better how our query-based model compares to other models, we performed an additional human evaluation experiment. We used the following criteria: *coherence*, *non-redundancy*, and *fluency*. As previously, we used the Best-Worst scaling on the Amazon test set. We assigned three AMT workers to each tuple containing summaries from PASS, AdaQSum (5%) + L1O, LexRank, and human annotators (Gold).

The results in Table 6 suggest that summaries produced by our model are more fluent and non-redundant than the ones produced by PASS. In general, PASS produces more diverse and detailed summaries yet with more semantic mistakes that make them harder to understand (hence lower fluency scores). However, summaries by both systems are similarly preferred in terms of coherence. Also, we note that PASS utilizes a separately trained classifier on human-annotated summaries (Fabbri et al., 2021) to rank candidate summaries, while our approach does not.

Input Content Fidelity. As was shown in Falke et al. (2019); Tay et al. (2019), the ROUGE metric can be insensitive to hallucinations (Maynez et al., 2020). However, hallucinations can lead to user aversion, and their reduction remains an open problem in summarization. To assess the input fidelity of generated summaries, we performed a human evaluation. Specifically, we used summaries produced by the adapter models (AdaSum (5%) + L1O and AdaQSum (5%) + L1O), FewSum, PASS, and human-written (Gold). In each task (HIT), we presented both reviews and all summary sentences. We asked three workers to assess how well the content in summary sentences is supported by the reviews. The three following options were available. *Full support*: all the content is reflected in the reviews; *Partial support*: only some content is reflected in the reviews; *No support*: content is not reflected in the reviews. The results, normalized by sentences, are shown in Table 7.

First, we observe that FewSum hallucinates the most, potentially because it was not initialized with a pre-trained language model. Second, PASS improves input fidelity over FewSum yet substantially underperforms our adapter-based models. We also notice a slight decrease in input fidelity when the query is used. This is likely caused by more abstractive summaries generated by AdaQSum, we discuss it in Sec. 5.3.

5 Analysis

5.1 Query-based Pre-training

Query-based summarizers should generate summaries reflecting all aspects in user queries to be useful in practice. We investigated how summarizers learn this task in the few-shot regime with and without pre-training. We created test-time queries from gold summaries (indicated by *) and input reviews. Further, we calculated the aspect recall (AR) score by counting aspect keywords in queries present in generated summaries. The results are shown in Table 8.

As indicated by low AR scores, without pre-training, the models miss many aspects in queries. The increase to nearly 100% in AR suggests that pre-training is crucial for the task. The same trend remains when aspect keywords from reviews are used in queries.

	R1	R2	RL	AR
FULL (100%) + Q*	40.52	10.96	25.06	59.84
FULL (100%) + L1O + Q*	42.65	11.53	26.82	96.39
ADAQSUM (5%)*	41.04	11.08	25.46	60.64
ADAQSUM (5%) + L1O*	43.84	13.41	27.31	97.19
ADAQSUM (5%)	38.58	10.10	24.19	69.14
ADAQSUM (5%) + L1O	38.53	10.52	26.06	98.78

Table 8: Amazon test set ROUGE F1 for query-based summarization. Here, * indicates that queries were created from gold summaries; AR stands for aspect recall.

	PPL↓
FULL (100%) + L1O	21.51
FULL (100%) + L1O + FT	34.87 (+13.36)
ADASUM (5%) + L1O	19.69
ADASUM (5%) + L1O + FT	**28.45** (**+8.76**)

Table 9: Catastrophic forgetting evaluation on the Amazon pre-training task's validation set, before and after fine-tuning (FT).

5.2 Catastrophic Forgetting

ROUGE scores in Table 4 suggest that L1O pre-training is beneficial for the end task. However, fine-tuning on summaries can lead to the catastrophic forgetting of the acquired in-domain specifics from reviews. Because adapters have fewer parameters to optimize, we hypothesized that they might be more robust to this phenomenon.

To test the hypothesis, we evaluated two models on the pre-training L1O pairs where a review is used as a summary, before and after fine-tuning on human-written summaries. For the first model, we optimized only adapters (5%), both in pre-training and fine-tuning. And in the second case, we optimized the entire model. We used PPL to measure the model's confusion about the pre-training pseudo summaries, as shown in Table 9.

The results demonstrate that the adapter-based model better preserves information about reviews after they are fine-tuned on summaries, as indicated by lower PPL scores. Our findings are also supported by Yu et al. (2021).

5.3 Abstractiveness

Abstracting information in reviews is important for practical applications (Carenini and Cheung, 2008). To investigate how well the models abstract, we computed the number of novel n-grams in generated summaries with respect to input reviews on the Amazon test set. The results in percentages are shown in Table 10.

First, we observe that FEWSUM tends to produce the most abstractive summaries, followed by PASS. Second, ADAQSUM has higher abstractivness than ADASUM. We also observe that abstractiveness is inversely proportional to input faithfulness in Table 7, in line with previous studies (Durmus et al., 2020; Dreyer et al., 2021).

	2-gram	3-gram	4-gram
FEWSUM	**78.63**	**95.59**	**98.74**
PASS	70.72	86.32	93.24
ADASUM (5%) + L1O	55.47	78.24	86.78
ADAQSUM (5%) + L1O	56.27	79.18	88.48

Table 10: The abstractiveness of generated summaries in terms of novel n-grams on the Amazon test set.

5.4 Semantic Mistakes

When a pre-trained model (with and without adapters) is fine-tuned on a handful of annotated samples, it often results in summaries with subtle semantic mistakes; see examples in Table 11. For instance, a 5% adapter model generates a semantically contradicting fragment: *'This **dead on arrival battery** is of good quality and holds a charge well.'*.

We hypothesize that it is caused by the lack of in-domain knowledge, which we propose to learn via L1O (see Sec. 2.4). During a manual investigation, we observed that L1O pre-training substantially reduces semantic mistakes. This is also reflected in higher ROUGE scores in Table 4.

6 Related Work

Extractive opinion summarization has been an active area of research (Hu and Liu, 2004; Ganesan et al., 2010; Medhat et al., 2014; Isonuma et al., 2019; Angelidis et al., 2020). For example, a more recent extractive method of Angelidis and Lapata (2018) decouples the summarization procedure into multiple steps with separate models. Other earlier approaches (Gerani et al., 2014; Di Fabbrizio et al., 2014) relied on text planners and templates, which, however, restrict the output.

Abstractive opinion summarization is an emerging branch (Chu and Liu, 2019; Amplayo and Lapata, 2020; Bražinskas et al., 2020b, 2021a). Customer reviews were used to train unsupervised summarizers in Amplayo and Lapata (2020); Bražinskas et al. (2020b); Isonuma et al. (2021). The few-shot model FEWSUM (Bražinskas et al., 2020a) was also pre-trained on customer reviews before fine-tuning. In this work instead, we focus on pre-training adapters to avoid catastrophic forgetting

AdaSum (5%)	This Thomas the Train costume is cute and functional. The size is perfect for a toddler or 2 year old and the candy pouch is large enough to carry a lot of candy. The costume can be customized with googly eyes, pumpkin patch, spiders, bats, or train tracks to make it more suitable for a child of any age. The fit is comfortable and the **fit quality is great**. The only complaint I have is **the quality of the felt fabric**.
AdaSum (5%) + L1O	This Thomas the Train costume is very cute and the size is perfect for a 2-year-old. The hat is thin and flimsy and the face is not 3D sculpted. The candy pouch is a nice feature and it looks like it will grow with the child. Overall, it's a cute costume and will be used for Halloween next year.
AdaSum (5%)	**This dead on arrival battery is of good quality and holds a charge well**. It is easy to install and is a great value for the money. **However, it may not hold a charge as advertised due to the plastic case bulging**. Overall, this product is highly recommended.
AdaSum (5%) + L1O	This battery is a great value for the price and works great. It is a good quality battery that can be used to replace a dead battery in an alarm system. The price is great and the quality of the product is good. The shipping was fast and the customer service was excellent.

Table 11: Adapter-based models (5%) and their generated outputs with and without L1O pre-training. Semantic mistakes and disfluencies are highlighted in **bold**.

and reduce computational and memory overheads. OpinionDigest (Suhara et al., 2020) proposes to aggregate opinions in a pipeline framework. We approach the problem end-to-end and rely on aspect keywords (e.g., price) instead of opinion phrases (e.g., good location). Controllability using input fragments (e.g., entities) and meta information (e.g., coarse-grained aspects) has received recent attention in various NLP domains (Fremann and Klementiev, 2019; Liu and Chen, 2021; Narayan et al., 2021; Elsahar et al., 2021). In contrast to a related work, SelfSum (Elsahar et al., 2021), we use aspect keywords instead of generic tokens and consider a few-shot setup instead of unsupervised and a different model architecture. Also, planning was tackled in opinion summarizeration in Amplayo et al. (2021b). However, their approach is substantially less flexible, as the summary plan consists of an aspect and sentiment classes only. Query-based settings have received recent attention in the news domain (Xu and Lapata, 2020, 2021). Compared to a concurrent work on opinion summarization AceSum (Amplayo et al., 2021a), our approach does not require a trained aspect induction model, is few-shot instead of self-supervised, and benefits from a large collection of automatically created fine-grained aspects (a couple of thousands) instead of human annotated coarse-grained aspects (up to 18). Concurrently with our work, Poth et al. (2021) support our findings on the benefits of pre-training adapters for other tasks.

7 Conclusions

In this work, we improve few-shot learning for opinion summarization with adapters pre-training on customer reviews in the end task-specific manner. In this way, the model learns in-domain specifics, which reduces semantic mistakes in generated summaries. We show that our approach leads to more than 2.0 and 1.3 ROUGE-L points improvement over the entire model's fine-tuning on the Amazon and Yelp datasets, respectively. Further, we propose a simple method for few-shot query-based summarization. The queries consist of aspect keywords reflecting potential user interests. We create these queries automatically and show that pre-training is crucial for the end task, significantly improving performance. Finally, in human evaluation, we demonstrate that the query-based model generates more coherent and less redundant summaries.

Acknowledgments

We would like to thank Jonathan Pilault, Leonardo Ribeiro, Sandeep Atluri, and the Amazon AI teams for their useful feedback and discussions.

8 Ethics Statement

We used only publicly available datasets. For human evaluation, we used a publicly available service (Amazon Mechanical Turk) to hire voluntary participants, requesting native speakers of English. The participants were fairly compensated, above the minimum hourly wage in their self-identified countries of residence.

9 Limitations

In this work, we explicitly focus on multi-document abstractive opinion summarization. However, our pre-training self-supervised method and fine-tuning techniques can be applied to a broader set of multi-document summarization domains (e.g., news) and can be considered in the future work. Also, while we tested our approach only with BART, we believe that it would work with other pre-trained encoder-decoder models, like PEGASUS (Zhang et al., 2020).

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021a. Aspect-controllable opinion summarization. *arXiv preprint arXiv:2109.03171*.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021b. Unsupervised opinion summarization with content planning. In *AAAI*.

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2020. Extractive opinion summarization in quantized transformer spaces. In *In Transactions of the Association for Computational Linguistics (TACL)*.

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.

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021a. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arthur Bražinskas, Mengwen Liu, Ramesh Nallapati, Sujith Ravi, and Markus Dreyer. 2021b. Transductive learning for abstractive news summarization. *arXiv preprint arXiv:2104.09500*.

Giuseppe Carenini and Jackie C. K. Cheung. 2008. Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1223–1232.

Giuseppe Di Fabbrizio, Amanda Stent, and Robert Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 54–63.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2021. Analyzing the abstractiveness-factuality tradeoff with nonlinear abstractiveness constraints. *CoRR*, abs/2108.02859.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Hady Elsahar, Maximin Coavoux, Jos Rozen, and Matthias Gallé. 2021. Self-supervised and controlled multi-document opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1646–1662, Online. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

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.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. 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 2208–2222, Online. Association for Computational Linguistics.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. 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)*, pages 591–602, Hong Kong, China. Association for Computational Linguistics.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2021. Unsupervised abstractive opinion summarization by generating sentences with tree-structured topic guidance. *Transactions of the Association for Computational Linguistics*, 9:945–961.

Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2142–2152, Florence, Italy. Association for Computational Linguistics.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2017. Measuring catastrophic forgetting in neural networks. *arXiv preprint arXiv:1708.02072*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. 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. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta

Cana, Dominican Republic. Association for Computational Linguistics.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *arXiv preprint arXiv:2106.04647*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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 2267–2277. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, 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, Yao Zhao, Joshua Maynez, Gonçalo Simoes, and Ryan McDonald. 2021. Planning with entity chains for abstractive summarization. *arXiv preprint arXiv:2104.07606*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and hb. *Sequim, WA: Sawtooth Software*.

Nadav Oved and Ran Levy. 2021. Pass: Perturb-and-select summarizer for product reviews. In *ACL/IJCNLP*.

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

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? efficient intermediate task selection. *arXiv preprint arXiv:2104.08247*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

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.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798.

Wenyi Tay, Aditya Joshi, Xiuzhen Zhang, Sarvnaz Karimi, and Stephen Wan. 2019. Red-faced ROUGE: Examining the suitability of ROUGE for opinion summary evaluation. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 52–60, Sydney, Australia. Australasian Language Technology Association.

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.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP)*, pages 3632–3645.

Yumo Xu and Mirella Lapata. 2021. Text summarization with latent queries. *arXiv preprint arXiv:2106.00104*.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. Adaptsum: Towards low-resource domain adaptation for abstractive summarization. In *arXiv preprint arXiv:2103.11332*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of International Conference on Machine Learning (ICML)*.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92.

10 Appendices

10.1 Human Evaluation Setup

To performed the human evaluation experiments described in Sec. 4.2, we hired workers with 98% approval rate, 1000+ HITS, Location: USA and the maximum score on a qualification test that we had designed. The test asked if the workers were native English speakers, and verified that they correctly understood the instructions of both the best-worst scaling and content support tasks. We paid the workers an approximate amount of $12 per hour.

10.2 Best-Worst Scaling Details

We performed human evaluation based on the Amazon test set using the AMT platform. We assigned 3 workers to each tuple containing summaries from different systems. We showed summaries and asked to select the best and worst one based on the criterion presented below.

1. *Fluency*: the summary sentences should be grammatically correct, easy to read and understand;

2. *Coherence*: the summary should be well structured and well organized;

3. *Non-redundancy*: there should be no unnecessary repetition in the summary.

Gold	These transition tights are perfect for children sensitive to the tight sensation other tights have around the foot. The material is soft and durable; they stand up well to both the rough nature of children, and the washing machine. This product does tend to run slightly small, so purchasing one size up is recommended.
FewSum	These tights are a great value for the price. The fit is true to size and the quality of the tights is very good. They are well made and will last a long time. They do run a little on the small side, so order a size up.
PASS	These soft, breathable tights are great for transitioning from tap to ballet. They fit snugly around the body and stay in place when worn with ballet shoes. They are well made and well made, and can last longer than other tights available. The colors are beautiful and will definitely be purchasing again.
AdaSum (5%) + L1O	These tights are soft and comfortable and fit well. They are durable and will last a long time. They can be worn with sandals or flip-flops. They do run small and should be ordered one size up to avoid squishing toes. The color is beautiful and the material is soft and durable.
AdaQSum (5%) + L1O	These tights are soft, comfortable, and durable. The **color** is beautiful and the **fit** is perfect for **tap** and ballet. They **fit** well and are durable enough to **wear** with flip-flops to class. They are recommended to order one size up if your child is chubby or slim.
Review 1	These are the perfect tights for my 5-year old. The tights are very well made and have already lasted several washings (hang dry). The color is beautiful, and my daughter loves that she can wear flip-flops to class like the big girls do.
Review 2	my 3 year old fit into these perfectly. I love these tights, they are great for wearing sandals to dance class and then pulling them over her toes to put ballet slippers on. They are nice and soft and the pink color is pretty. Will purchase again.
Review 3	These are my daughters preferred ballet tights. They fit well and don't squish her toes as much as some others. The convertible option is nice as she can wear flip flops to the studio with her tights. I like that they appear to be fairly durable.
Review 4	Bought this for my little one to use for her ballet class. She's almost 4 and this fits perfectly. Transition tights give her the ability to pull up the foot area to around the ankles so that they don't get dirty when not wearing her shoes, but fit well and stay in place when pulled over her feet and used with ballet shoes on.
Review 5	Great soft fabric, runs small a though. U should consider getting one size up to avoid having your daughter, or son if he's into ballet, have little circulation. Don't get me wrong, great product and material.
Review 6	I purchased these tights for my 4 year old because she has a tap class immediately followed by an acro class. Tights fit well and were easy to transition to bare feet after tap. I can't comment on how they hold up, as we only had one week of dance so far.
Review 7	This is my go to tight when my daughter needs new ones. I won't buy anything else. They are super comfy and last forever! She is 7 but very slim and these will fit her for a while but the 8-12 is a better fit for longevity purposes but they only had this size left and I needed them right away.
Review 8	These are great tights. If your girl is chubby on the waist i think you should order a number up. they are a strong tihgts. My daugther has it for more than 4 months and hasn't broken them. She only complains some days about the waist, just when she has had a lot of milk or food recently. she's 3 years old.

Table 12: Example summaries produced by different systems for an Amazon product. We highlight in **bold** keywords that were part of the query for AdaQSum (5%) + L1O.

GOLD	The results with this scanner are sporadic at best- there are many problems with the software drivers and the stringent system requirements that come along with it. It does not work with the most recent versions of Windows, and the company charges for tech support calls.
FEWSUM	This is a great scanner. It is easy to use and easy to set up. The software that comes with it does not work with Windows Vista. It would be nice if it was a little more user friendly, but for the price it can't be beat.
PASS	The VuPoint Digital Scanner is very user friendly and easy to use. It can scan over 1200 35 mm negatives and about a dozen slides. It takes time and money to get the scanner to fit in the holders that come with it. The company also charges $10.00 per phone call for tech support.
ADASUM (5%) + L1O	The VuPoint Digital Scanner is easy to use and does a great job converting negatives to digital format. However, the software is not compatible with newer versions of Windows. The company ArcSoft charges $10 per phone call for tech support. Overall, this product is not recommended.
ADAQSUM (5%) + L1O	The VuPoint Digital Scanner does a great job of converting **photo** negatives to digital format. The **software** is easy to use and easy to install. However, the **image** bleaches out with too much light. A **replacement** unit is required. Overall, this product is recommended.
REVIEW 1	I recently bought this film and slide scanner to scan my grandfather's slide collection. It bleaches out the image with too much light. I tried changing the settings to improve the image quality, but had no luck. The company ArcSoft charges $10.00 per phone call for tech support. You are better off making the investment on a nicer quality scanner.
REVIEW 2	* * Not Reccommended * * Purchased as a gift in August. Opened a week ago. Spent the last week trying to get Win Xp to recognise the Vu Point scanner. Many drivers and reloads later all I have is a little black box with a red light and a message from windows that says 'USB Device Not Recognised'.
REVIEW 3	I used the VuPoint Digital Scanner to scan over 1,200 35 mm negatives and about a dozen slides and found this gadget a most user-friendly and efficient tool. I even managed to upload a few black and white negatives from 1963. I recommend the product highly.
REVIEW 4	While the software was good for Windows XP and Vista, I now have Windows 7 and would like to have software for the newer operating system. The company prefers to sell other products rather than update their software. I can't see recommending this product in today's market.
REVIEW 5	While most equipment will work with more modern versions of Windows than were available when manufactured this is not true with this scanner. Requires Windows XP means it won't work with earlier OR LATER. Its on its way back for a refund.
REVIEW 6	I found the VuPoint scanner not acceptable and I am still waiting for a replacement. My contacts with VuPoint were helpful but the equipment still did not produce acceptable images. My contact with the seller has been sporadic, at best, and a replacement unit has not been delivered.I an NOT anxious to deal with these providers again.
REVIEW 7	Product is very easy to use. Does a great job converting my slide and photo negatives to digital format. Touch-up and enhance program gave me just what I needed to clean up and enhance some of the scans, Company was great to work with!!
REVIEW 8	Its not worth the time it takes to get the negative to fit in the holders they give you. I'd much rather buy a hp flat bed scanner that lets you see the final photo image and not just an image of the negative. It takes to much time and isn't worth the money.

Table 13: Example summaries produced by different systems for an Amazon product. We highlight in **bold** keywords that were part of the query for ADAQSUM (5%) + L1O.