Aspect-Controllable Opinion Summarization

Reinald Kim Amplayo **Stefanos Angelidis** **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, EH8 9AB
reinald.kim@ed.ac.uk s.angelidis@ed.ac.uk mlap@inf.ed.ac.uk

Abstract

Recent work on opinion summarization produces general summaries based on a set of input reviews and the popularity of opinions expressed in them. In this paper, we propose an approach that allows the generation of customized summaries based on aspect queries (e.g., describing the location and room of a hotel). Using a review corpus, we create a synthetic training dataset of (review, summary) pairs enriched with *aspect controllers* which are induced by a multi-instance learning model that predicts the aspects of a document at different levels of granularity. We fine-tune a pre-trained model using our synthetic dataset and generate aspect-specific summaries by modifying the aspect controllers. Experiments on two benchmarks show that our model outperforms the previous state of the art and generates personalized summaries by controlling the number of aspects discussed in them.

General
The room was clean and comfortable. The staff was very friendly and helpful. It was a great location, just a short walk to the beach. There wasn't much to do in the area, but the food was good.
Location
The location was great, right on the Boardwalk, and close to the Venice beach.
Rooms
The room was very clean and the bathroom was very nice. The bathroom had a large separate shower. There was a TV in the room.
Location and **Rooms**
The location is great, right on Boardwalk, and the beach is very nice. The room was very clean and the bathroom was very nice and the shower was great.
Cleanliness, **Location**, **Room**, and **Service**
The staff was very friendly and helpful. The room was very clean, and the bathroom was very nice. It was a great location, right on the beach.

Table 1: General and aspect-specific summaries generated by our model for a hotel from the SPACE dataset. Aspects and aspect-specific sentences are color-coded.

1 Introduction

Consumers oftentimes resort to review websites to inform their decision making (e.g., whether to buy a product or use a service). The proliferation of online reviews has accelerated research on opinion mining (Pang and Lee, 2008), where the ultimate goal is to glean information from reviews so that users can make decisions more efficiently. Opinion mining has assumed several guises in the literature such as sentiment analysis (Pang et al., 2002), aspect extraction (Hu and Liu, 2004; He et al., 2017), combinations thereof (Mukherjee and Liu, 2012; Pontiki et al., 2016), and notably *opinion summarization* (Hu and Liu, 2006; Wang and Ling, 2016), whose aim is to create a textual summary of opinions found in multiple reviews.

Text summarization models, both extractive (Narayan et al., 2018; Zheng and Lapata, 2019; Cachola et al., 2020) and abstractive (See et al., 2017; Gehrmann et al., 2018; Liu and Lapata, 2019), operate under the assumption that salient content is relevant (Erkan and Radev, 2004) and should be presented in the summary. Opinion summarization is no exception, focusing on creating summaries based on opinions that are *popular* or *redundant* across reviews (Angelidis and Lapata, 2018b; Chu and Liu, 2019; Amplayo and Lapata, 2020; Bražinskas et al., 2020; Amplayo et al., 2021).

However, the notion of salience in reviews largely depends on *user interest*. For example, one might only care about the connectivity of a television product, an aspect which might be unpopular amongst reviews. As a result, models that create *general* opinion summaries may not satisfy the needs of all users, limiting their ability to make decisions. Angelidis et al. (2021) mitigate this problem with an extractive approach that produces both general and *aspect-specific* opinion summaries. They achieve this essentially by clustering opinions through a discrete latent variable model (van den Oord et al., 2017) and extracting sentences based on popular aspects or a particular

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6578–6593
November 7–11, 2021.

aspect. By virtue of being extractive, their summaries can be incoherent, and verbose containing unnecessary redundancy. And although their model creates summaries for individual aspects, it is not clear how to control the number of aspects in the output (e.g., to obtain summaries that mention multiple rather than a single aspect of an entity).

In this paper, we propose an abstractive opinion summarization model that generates aspect-controllable summaries. Using a corpus of reviews on entities (e.g., hotels, television sets), we construct a synthetic training dataset consisting of reviews, a pseudo-summary, and three types of *aspect controllers* which reflect different levels of granularity: aspect-related keywords, review sentences, and document-level aspect codes. We induce aspect controllers automatically based on a multiple instance learning model (Keeler and Rumelhart, 1991) and very little human involvement. Using the aspect-enriched dataset, we then fine-tune a pretrained model (Raffel et al., 2020) on summary generation. By modifying the controllers, we can flexibly generate general and aspect-specific summaries, discussing one or more aspects. Figure 1 shows summaries generated by our model.

We perform experiments on SPACE (Angelidis et al., 2021), a single domain dataset consisting of hotel reviews, and OPOSUM (Angelidis and Lapata, 2018b), a dataset with product reviews from multiple domains (e.g., "laptop bags", "boots"). Automatic and human evaluation show that our model outperforms previous approaches on both tasks of general and aspect-specific summarization. We also demonstrate that it can effectively generate multi-aspect summaries based on user preferences. We make our code and data publicly available.[1]

2 Related Work

Earlier work on opinion summarization has focused on general summarization using extractive (Hu and Liu, 2006; Kim et al., 2011; Angelidis and Lapata, 2018b) or abstractive methods (Ganesan et al., 2010; Carenini et al., 2013; Fabbrizio et al., 2014). Due to the absence of opinion summaries in review websites and the difficulty of annotating them on a large scale, more recent methods consider an unsupervised learning setting where there are only reviews available without corresponding summaries (Chu and Liu, 2019; Bražinskas et al., 2020). They make use of autoencoders (Kingma and Welling, 2014) and variants thereof to learn a review decoder through reconstruction, and use it to generate summaries conditioned on averaged representations of the inputs.

A more successful approach to opinion summarization is through the creation of synthetic datasets, where (review, summary) pairs are constructed from a review corpus to enable supervised training. These methods usually start by randomly selecting a review which they treat as a pseudo-summary and subsequently pair it with a set of input reviews based on different strategies. These include random sampling (Bražinskas et al., 2020), generating noisy versions of the pseudo-summary (Amplayo and Lapata, 2020), ranking reviews based on similarity and relevance (Elsahar et al., 2021), and making use of content plans to create more naturalistic pairs (Amplayo et al., 2021).

Our work is closest to Angelidis et al. (2021) who propose an extractive summarization model that uses a vector-quantized variational autoencoder (van den Oord et al., 2017) to learn aspect-specific review representations. Their model effectively groups opinion sentences into clusters and extracts those capturing aspect-relevant information. We employ multi-instance learning to identify aspect-bearing elements in reviews with varying degrees of granularity (e.g., words, sentences, documents) which we argue affords greater flexibility and better control of the output summaries. In doing so, we also introduce an effective method to create synthetic datasets for aspect-guided opinion summarization. Our work also relates to approaches which attempt to control summarization output based on length (Kikuchi et al., 2016), content (Fan et al., 2018), style (Cao and Wang, 2021), or textual queries (Dang, 2006). Although we focus solely on aspect, our method is general and could be used to adjust additional properties of a summary such as sentiment (e.g., positive vs. negative) or style (e.g., formal vs. colloquial).

3 Problem Formulation

Let C denote a corpus of reviews about entities (e.g., products, hotels). Let $R_e = \{r_1, r_2, ..., r_N\}$ denote a set of reviews for entity e and $A_e = \{a_1, a_2, ..., a_M\}$ a set of aspects that are relevant for the entity (e.g., cleanliness, location). Each review r_i is a sequence of tokens $\{w_1, w_2, ...\}$, while each aspect a_j is represented by a small set of *seed words* $\{v_1, v_2, ...\}$ (e.g., *spotless*, *dirty*, *stain*).

[1] https://github.com/rktamplayo/AceSum

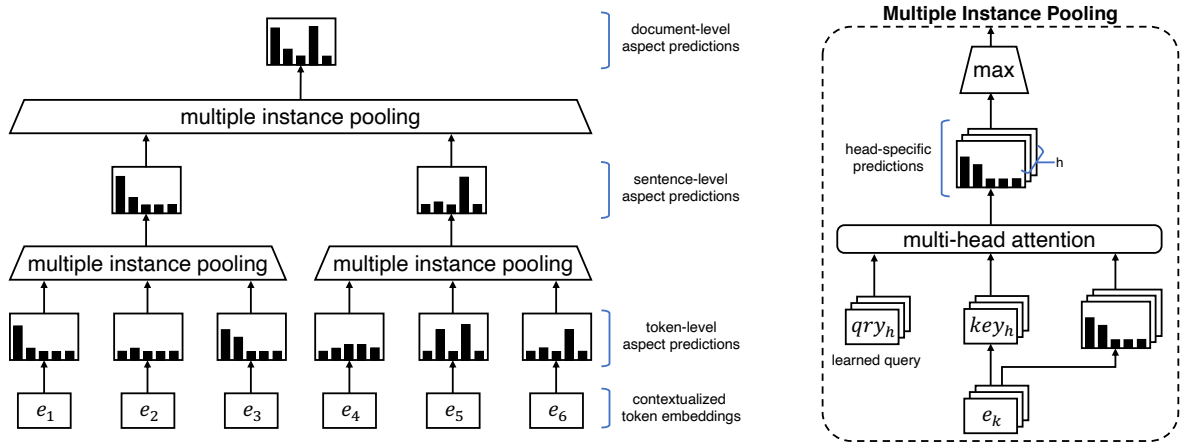

Figure 1: Overview of the controller induction model. Token-level aspect predictions are aggregated into sentence-level predictions using a multiple instance pooling mechanism (described on the right). The process is repeated from sentence- to document-level predictions.

These seed words can be acquired automatically (Angelidis and Lapata, 2018b) or provided by users (see Appendix for those used in our experiments).

Our approach creates two types of summaries: (a) a general summary that contains salient opinions about *all* aspects of an entity, and (b) an aspect-specific summary that focuses on opinions about *particular* aspects of interest specified by a query $Q = \{q_1, q_2, ..., q_M\}$; here, q_j is an indicator function which designates whether the aspect should be mentioned in the summary. We emphasize that the query can represent more than one aspect to reflect real-world usage. To facilitate supervised training, we create a synthetic training dataset $D = (X, z, y)$, which is a set of triples composed of input reviews X, a pseudo-summary y, and aspect controllers z (Section 3.2). Our aspect controllers are induced with a unified model based on multi-instance learning (Section 3.1) and correspond to different levels of granularity: (1) document-level aspect codes, (2) aspect-related review sentences, and (3) aspect keywords.

At training time, we fine-tune a pretrained sequence-to-sequence Transformer model (Raffel et al., 2020) using controllers z as input and a pseudo-summary as output. During inference, we modulate summary generation by modifying the controllers, e.g., we produce a general summary using all aspect codes, or an aspect-specific one based on a subset thereof (Section 3.3).

3.1 Controller Induction Model

A key feature of our approach is the set of aspect controllers which allow our summarization model to be controllable. We induce these controllers using a multiple instance learning (MIL) model, illustrated in Figure 1. MIL is a machine learning framework where labels are associated with groups of instances (i.e., *bags*), while instance labels are unobserved (Keeler and Rumelhart, 1991). The goal is then to infer labels for bags (Dietterich et al., 1997; Maron and Ratan, 1998) or jointly for instances and bags (Zhou et al., 2009; Wei et al., 2014; Kotzias et al., 2015; Xu and Lapata, 2019; Angelidis and Lapata, 2018a). Our MIL model is an example of the latter variant.

In our setting, documents are bags of sentences and sentences are bags of tokens. We further assume that only documents have aspect labels. Given review r with tokens $\{w_k\}$, we obtain token encodings $\mathbf{e} = \{e_k\}$ from a pretrained language model (PLM; Liu et al. 2019) which uses the popular Transformer architecture (Vaswani et al., 2017). We use a non-linear transformation to obtain token-level aspect predictions $\mathbf{z}_\mathcal{T}$:

$$\mathbf{e} = \text{PLM}(\{w_k\}) \quad (1)$$

$$\mathbf{z}_\mathcal{T} = \tanh(W\mathbf{e} + b) \quad (2)$$

where $\mathbf{z}_\mathcal{T} \in \mathbb{R}^{N \times M}$, and N and M are the number of tokens and aspects, respectively. A positive value denotes that the token is related to the aspect of interest (and otherwise unrelated).

Multiple Instance Pooling To obtain sentence-level aspect predictions $\mathbf{z}_\mathcal{S}$, we aggregate token-level predictions $\mathbf{z}_\mathcal{T}$ using a new pooling method particularly effective for our multi-instance learning setting. We first obtain multiple predictions $\mathbf{z}_h$

for each attention head h:

$$\mathbf{z}_h = \sum\nolimits_k (\mathbf{z}_\mathcal{T} * a_h[k]) \quad (3)$$

$$a_h = \text{softmax}(key_h \cdot qry_h) \quad (4)$$

where $*$ is element-wise multiplication, $\cdot$ is dot product, k is the token index, qry_h is a head-specific query vector, and key_h is defined below:

$$key_h = \tanh(W_h \mathbf{e} + b_h) \quad (5)$$

We hypothesize that different attention heads represent different aspects of the semantic space, and are thus helpful at predicting multiple aspects. We obtain a sentence-level prediction by max pooling the predictions of individual heads:

$$\mathbf{z}_S = \text{max-pool}(\{\mathbf{z}_h\}) \quad (6)$$

We use max pooling since we want to isolate the most pertinent aspects for a given sentence; standard pooling methods such as mean and attention pooling (Angelidis and Lapata, 2018a; Xu and Lapata, 2019) assume that *all* instances of a bag contribute to its label. In Figure 1 (right) we illustrate our pooling mechanism and empirically show in experiments (see Section 5.1) it is superior to alternatives.

We so far discussed how multiple instance pooling is applied at the token-level to obtain sentence-level predictions $\mathbf{z}_\mathcal{S}$. Analogously, multiple instance pooling is applied to sentences to obtain document-level predictions $\mathbf{z}_\mathcal{D}$ (see Figure 1).

Training and Inference Training the multiple instance model just described requires a dataset consisting of (review, aspect label) pairs. Unfortunately, we do not have access to annotations denoting which aspects are discussed in each review. Recall, however, that aspects are represented by seed words $\{v_1, v_2, ...\}$, which we exploit to induce silver-standard labels. Specifically, for each review in the dataset, we obtain binary labels $\hat{\mathbf{z}}_\mathcal{D}$ where $\hat{\mathbf{z}}_\mathcal{D}[a] = 1$ if at least one seed word for aspect a is found in the review (and -1 otherwise).

We train the model using a soft margin loss, summing over all aspects $a \in A$:

$$\mathcal{L}_{ctrl} = \sum\nolimits_a \log(1 + \exp(-\mathbf{z}_\mathcal{D}[a] * \hat{\mathbf{z}}_\mathcal{D}[a])) \quad (7)$$

The parameters of the pretrained language model (see Equation (2)) are frozen, i.e., they are not fine-tuned during training which makes our controller induction model lightweight and efficient.

Summary y: At first they took us to an unready room which was disappointing but after a short wait they took us to a really big room with a great harbor scene as an apology to the mess. The rooms are pretty new or renovated recently. Bathroom is clean and wide. The beds are comfortable and big.

Review x_1: Check in was quick and our bags were brought to the room in a timely manner. The rooms and hallways left a little more to be desired. The rooms didnt look nearly as good as they did in other less known cities. No safe or frig in the rooms. The staff was great.

Review x_2: Only option for a hot meal for breakfast was scrambled eggs and bacon; The toaster was broken as well, with burned out elements. Other food in the lounge was good (fruit, coffee). Recommendation: eat elsewhere; even room service would probably have been better.

Figure 2: Pseudo-summary y and input reviews X; the aspect code for summary y is room. Review sentences with the same aspect are underlined and same aspect-keywords are magnified.

3.2 Synthetic Dataset Creation

The MIL model allows us to learn three kinds of aspect controllers which are subsequently used to create a synthetic dataset for training our summarizer. These are *aspect codes*, essentially document-level aspect predictions $\mathbf{z}_D$, which control the overall aspect of the summary, *aspect keywords* ensure content support by explicitly highlighting which tokens from the input should appear in the summary, and *aspect-relevant sentences* which provide textual context for summary generation (while non-aspect-related sentences are ignored).

We first sample review r_i as a pseudo-summary from review set R_e of entity e. We treat r_i as a pseudo-summary provided it resembles a real summary. We assume that opinion summaries discuss specific aspects regarding entity e. We use our controller induction model to verify this, i.e., document-level aspect predictions $\mathbf{z}_\mathcal{D}$ for r_i should be positive for at least one aspect. Provided r_i fulfills this constraint, we use it as summary y and $R_e - \{r_i\}$ as review set X. A simplified example is shown in Figure 2, the pseudo summary is highlighted in gray and the input reviews in cyan. The summary focuses on the room aspect of a hotel and this is its aspect code (shown in blue).

Let (X, y) denote review set X for summary y (we only show two reviews in Figure 2 but there are usually hundreds). We obtain (positive) document-level aspect predictions $\mathbf{z}_D^{(y)}$ for summary y and sentence-level aspect predictions $\mathbf{z}_S^{(x)}$ for all re-

views $x \in X$. We then rank review sentences in X based on their similarity to the summary's overall aspect. Specifically, we compare predictions $\mathbf{z}_{\mathcal{S}}^{(x)}$ with $\mathbf{z}_{\mathcal{D}}^{(y)}$ using the soft margin loss function from Equation (7). We also compare token-level predictions $\mathbf{z}_{\mathcal{T}}^{(x)}$ with $\mathbf{z}_{\mathcal{D}}^{(y)}$ using the same function to induce aspect keywords. In Figure 2 sentences which discuss the same aspect as the summary are underlined, and same-aspect keywords are magnified. For illustration purposes we only show one aspect code in Figure 2, but these can be several, and different review sentences and keywords would be selected for different aspects.

3.3 Opinion Summarization Model

We use a pretrained sequence-to-sequence Transformer model (Raffel et al., 2020) to generate opinion summaries. We transform the aspect controllers z into the following format:

`[CODE]` `[ASPECT`$_2$`]` `[ASPECT`$_3$`]`
`[KEY]` keyword$_1$ keyword$_2$ ... `[SNT]`
first sentence `[SNT]` second sentence ...

where `[CODE]`, `[KEY]`, and `[SNT]` are indicators denoting that the next tokens correspond to aspect codes, keywords, and review sentences.

Instead of the full set of input reviews X, the encoder takes z as input and produces multi-layer encodings $\mathbf{Z}$. The decoder then outputs a token distribution $p(y_t)$ for each time step t, conditioned on both $\mathbf{Z}$ and $y_{1:t-1}$ through attention:

$$\mathbf{Z} = \text{Encoder}(z) \tag{8}$$

$$p(y_t) = \text{Decoder}(y_{1:t-1}, \mathbf{Z}) \tag{9}$$

We fine-tune the model using a maximum likelihood loss to optimize the probability distribution $p(y)$ based on gold summary $\hat{y}$:

$$\mathcal{L}_{gen} = -\sum\nolimits_t \hat{y}_t \log p(y_t) \tag{10}$$

During inference, we can generate different kinds of opinion summaries by modifying the aspect controllers. When creating a general summary, we use all aspect codes as input. Analogously, when generating a single aspect summary, we use one aspect code. The aspect codes guide the selection of keywords and sentences from the input reviews (see Figure 2) which are given as input to our Transformer model to generate the summary (see Equation (8)).

Dataset	SPACE	OPOSUM+
review corpus size	1.14M	4.13M
#domains	1	6
#aspects	6	18
#test examples	50	60
#reviews/example	100	10
#summaries/example	3	3
#general summaries	150	180
#aspect summaries	900	540

Table 2: Statistics for SPACE and OPOSUM++ (underlined summaries are extractive).

4 Experimental Setup

Datasets We performed experiments on two opinion summarization datasets covering different review domains. SPACE (Angelidis et al., 2021) is a large corpus of "hotel" reviews from TripAdvisor; it contains human-written abstractive opinion summaries for evaluation only. Each instance in the evaluation set consists of 100 reviews and seven summaries: one general summary and six aspect-specific ones representing the aspects building, cleanliness, food, location, rooms, and service. OPOSUM (Angelidis and Lapata, 2018b) is a large corpus of product reviews from six different domains: "laptop bags", "bluetooth headsets", "boots", "keyboards", "televisions", and "vacuums". It also includes an evaluation set with extractive general summaries. We extended this dataset by (a) adding aspect-specific summaries which are human-written and abstractive following the methodology from Angelidis et al. (2021), and (b) increasing the size of the corpus. We call this extended dataset OPOSUM+. Both datasets include five human-annotated seed words for each aspect (see Appendix for details). Data statistics are shown in Table 2. Using our synthetic dataset creation method, we were able to generate 512K and 341K training instances for SPACE and OPOSUM+, respectively.

Implementation For our pretrained Transformer models, we used weights and settings available in the HuggingFace library (Wolf et al., 2020). Specifically, we used `distilroberta-base` (Liu et al., 2019; Sanh et al., 2019) as our language model and `t5-small` (Raffel et al., 2020) as our sequence-to-sequence model. We trained the controller induction model with a learning rate of $1e{-}4$ for 100K steps, using $h = 12$ heads. For OPOSUM+, we trained separate controller induction models for different domains. For the aspect controllers, we selected 10-best keywords, and review

sentences were truncated up to 500 tokens to fit in the pretrained model. For summarization, we used a learning rate of $1e-6$ and 500K training steps. We used Adam with weight decay (Loshchilov and Hutter, 2019) to optimize both models. We added a linear learning rate warm-up for the first 10K steps. We generate summaries with beam search of size 2 and refrain from repeating ngrams of size 3 (Paulus et al., 2018).

5 Results

We compared our **A**spect **C**ontroll**e**d **Sum**marization (ACESUM) model with several extractive and abstractive approaches. Traditional extractive systems include selecting as a summary the review closest to the CENTROID (Radev et al., 2004) of the input reviews and LEXRANK (Erkan and Radev, 2004), a PageRank-like algorithm that selects the most salient sentences from the input. For both methods we used BERT encodings (Devlin et al., 2019) to represent sentences and documents. Other extractive systems include QT[2] (Angelidis et al., 2021), a neural clustering method that uses Vector-Quantized Variational Autoencoders (van den Oord et al., 2017) to represent opinions in quantized space, and ACESUMEXT, an extractive version of our model that uses sentences ranked by our controller induction model as input (truncated up to 500 tokens) to LexRank.

Abstractive systems include MEANSUM (Chu and Liu, 2019), an autoencoder that generates summaries by reconstructing the mean of review encodings, COPYCAT (Bražinskas et al., 2020), a hierarchical variational autoencoder which learns a latent code of the summary, and two variants of T5 (Raffel et al., 2020) trained with different synthetic dataset creation methods. For T5-RANDOM, summaries are randomly sampled (Bražinskas et al., 2020), whereas for T5-SIMILAR reviews are sampled based on their similarity to a candidate summary (Amplayo and Lapata, 2020).

Finally, we compared against two upper bounds: an extractive ORACLE which selects as a summary the review with the best ROUGE score against the input, and a HUMAN upper bound, calculated as inter-annotator ROUGE. Examples generated by our model are in Table 1 and the Appendix.

[2]We report results for QT using our seed words which are human-annotated. We also present results in the Appendix with their seed words which were automatically induced.

	SPACE			OPOSUM+		
Model	R1	R2	RL	R1	R2	RL
CENTROID	31.29	4.91	16.43	33.44	11.00	20.54
LEXRANK	31.41	5.05	18.12	35.42	10.22	20.92
QT	38.66	10.22	21.90	37.72	14.65	21.69
ACESUMEXT	35.50	7.82	20.09	**38.48***	**15.17***	**22.82***
MEANSUM	34.95	7.49	19.92	26.25	4.62	16.49
COPYCAT	36.66	8.87	20.90	27.98	5.79	17.07
T5-RANDOM	37.65	10.62	22.82	29.88	5.64	17.19
T5-SIMILAR	38.84	10.82	22.74	30.42	6.07	17.17
ACESUM	**40.37***	**11.51***	**23.23**	32.98	10.72	20.27
ORACLE	40.23	13.96	23.46	41.88	21.52	29.30
HUMAN	49.80	18.80	29.19	55.42	37.26	44.85

Table 3: Automatic evaluation for *general summarization*. Extractive/Abstractive/Upper-bound models are shown in first/second/third block. Best systems are boldfaced; an asterisk (*) means there is a significant difference between best and 2nd best systems (based on paired bootstrap resampling; $p < 0.05$).

5.1 Automatic Evaluation

We evaluated the quality of general and aspect-specific opinion summaries using F_1 ROUGE (Lin and Hovy, 2003). Unigram and bigram overlap (ROUGE-1/2) are proxies for assessing informativeness while the longest common subsequence (ROUGE-L) measures fluency.

General Opinion Summarization Table 3 reports results on general opinion summarization. As can be seen, ACESUM outperforms all competing models on SPACE and performs best among abstractive systems on OPOSUM+. Our extractive model, ACESUMEXT, is overall best on OPOSUM+. This is expected since general OPOSUM+ summaries are extractive. Amongst abstractive models, Transformer-based models outperform MEANSUM and COPYCAT, demonstrating that pretraining is helpful for opinion summarization.

Aspect-Specific Opinion Summarization Most comparison systems (all except QT) cannot naturally generate aspect-specific summaries. We use a simple sentence-filtering method to remove non-aspect-related sentences from the input during inference. Specifically, we use BERT encodings (Devlin et al., 2019) to represent tokens in review sentences $\{r_i^{(bert)}\}$ and aspect seeds $\{a_j^{(bert)}\}$. We then rank the review sentences based on the maximum similarity between seed and sentence tokens, calculated as $\max_{i,j}(\text{sim}(r_i^{(bert)}, a_j^{(bert)}))$, where $\text{sim}(a, b)$ is the cosine similarity function. This method cannot be ported to the CENTROID and ORACLE baselines, and thus we do not compare with them.

Our results are summarized in Table 4. Note

Model	SPACE R1	SPACE R2	SPACE RL	OPOSUM+ R1	OPOSUM+ R2	OPOSUM+ RL
LEXRANK	24.61	3.41	18.03	22.51	3.35	17.27
QT	28.95	8.34	21.77	23.99	4.36	16.61
ACESUMEXT	30.91	8.77	23.61	26.16	5.75	18.55
MEANSUM	25.68	4.61	18.44	24.63	3.47	17.53
COPYCAT	27.19	5.63	19.18	26.17	4.30	18.20
T5-RANDOM	21.40	4.83	15.45	24.47	4.20	16.18
T5-SIMILAR	22.69	5.12	16.44	23.86	4.30	16.36
ACESUM	**32.41***	**9.47***	**25.46***	**29.53***	**6.79***	**21.06***
HUMAN	44.86	18.45	34.58	43.03	16.16	31.53

Table 4: Automatic evaluation for *aspect-specific summarization*. Extractive/Abstractive/Upper-bound models are shown in first/second/third block. Best systems are boldfaced; an asterisk (*) means there is a significant difference between best and 2nd best systems (based on paired bootstrap resampling; $p < 0.05$).

Model	SPACE Doc F_1	SPACE Sent F_1	OPOSUM+ Doc F_1	OPOSUM+ Sent F_1
MIP (ours)	**77.35**	**40.85**	**83.28**	**50.48**
Max	63.35	35.12	66.52	44.00
Attention	31.77	29.30	34.00	35.80
Mean	27.38	27.87	30.38	34.35

Table 5: Performance of controller induction models (document- and sentence-level); comparison of multiple instance pooling (MIP) against max, mean, and attention pooling.

Model	SPACE General	SPACE Aspect	OPOSUM+ General	OPOSUM+ Aspect
ACESUM	23.23	25.03	19.64	20.16
No aspect code	22.29	24.99	17.22	17.54
No keywords	21.88	24.82	18.97	19.97
Random sentences	22.42	19.16	18.96	13.44

Table 6: Variants of ACESUM with different aspect controllers. Results are shown using ROUGE-L for general and aspect-specific opinion summaries.

that SPACE and OPOSUM+ focus exclusively on *single* aspect summaries. We assess our model's ability to generate summaries covering multiple aspects in the following section. Overall, ACESUM performs best across datasets and metrics, which shows that our controllers can effectively customize summaries based on aspect queries. Interestingly, amongst extractive models, ACESUMEXT performs best. This suggests that, a simple centrality-based extractive approach such as LexRank (Erkan and Radev, 2004) can produce good enough summaries as long as an effective sentence filtering method is applied beforehand (in our case this is based on the controller induction model). T5 models perform substantially worse on this task, indicating that synthetic datasets based on either random or similarity-based sampling techniques are not suited to aspect-specific opinion summarization.

Ablation Studies We present various ablation studies on the controller induction model and the summarization model itself. In Table 5, we compare our multiple instance pooling (MIP) mechanism with three standard pooling methods: mean, max, and attention-based pooling. We evaluate models using document and sentence F_1 which measures the quality of document- and sentence-level aspect predictions. We extrapolate aspect labels for documents and sentences from the development set which contains aspect-specific summaries. We assume the aspect for which a summary is written is the document label and that all sentences within the summary are also representative of the same aspect. Results show that attention and mean pooling are not suitable for multi-instance learning, underperforming especially on document-level F_1. This suggests that token-level predictions are not used effectively to predict higher level aspects. Our results confirm that using multiple experts (i.e., attention heads) yields better aspect predictions.

In Table 6, we evaluate the contribution of different aspect controllers to summarization output. Selecting sentences randomly rather than based on aspect hurts performance, in particular when generating aspect-specific summaries. We also find that aspect codes substantially increase model performance in OPOSUM+. We conjecture that this is due to OPOSUM+ having multiple domains and, consequently, more aspects compared to SPACE.

5.2 Human Evaluation

We conducted several human elicitation studies to further analyze the summaries produced by competing systems using the Amazon Mechanical Turk crowdsourcing platform.

Best-Worst Scaling The first study assessed the quality of general opinion summaries using Best-Worst Scaling (BWS; Louviere et al., 2015). Participants were shown a human-written summary, in relation to which they were asked to select the best and worst among system summaries, taking into account the following criteria: *Informativeness* (how consistent are the opinions with the reference?), *Coherence* (is the summary easy to read and well-organized?), *Conciseness* (does the summary provide useful information in a concise manner?), and *Fluency* (is the summary grammatical?).

We compared general summaries produced by

SPACE	Inf	Coh	Con	Flu
LEXRANK	−48.3	−38.4	−36.9	−43.3
T5-SIMILAR	5.8	11.2	17.2	0.6
QT	20.4	1.3	1.2	2.6
ACESUM	22.1	26.0*	18.5	38.8*

OPOSUM+	Inf	Coh	Con	Flu
LEXRANK	−27.3	−21.1	−18.2	−23.8
T5-SIMILAR	−31.1	10.0	4.7	−1.9
QT	20.3	−25.3	−21.6	−9.6
ACESUM	38.1*	36.3*	35.2*	35.3*

Table 7: *Best-Worst Scaling* evaluation. Best values are bold-faced. An asterisk (*) means that the system is significantly better than the second best system (one-way ANOVA with posthoc Tukey HSD tests, $p < 0.05$). Inf: informative, Coh: coherent, Con: concise, Flu: fluent.

SPACE	Exclusive	Partial	None
T5-SIMILAR	10.6	74.8	14.6
QT	43.8	39.0	17.1
ACESUM	50.9	42.6	6.5
HUMAN	64.9	31.6	3.5

OPOSUM+	Exclusive	Partial	None
T5-SIMILAR	9.4	48.2	42.5
QT	22.2	41.9	35.9
ACESUM	42.2	45.4	12.4
HUMAN	63.0	31.5	5.6

Table 8: Proportion of summaries that discuss the target aspect exclusively, partially, or not at all.

SPACE	All	One	Other	None
QT	10.0	35.3	34.7	20.0
ACESUM	61.3	19.3	18.0	1.3

OPOSUM+	All	One	Other	None
QT	18.8	27.5	33.6	20.1
ACESUM	47.0	16.8	26.8	9.4

Table 9: Proportion of target aspects discussed in system summaries (All: both aspects are mentioned; One: only one is mentioned; Other: other aspects are also mentioned; None: no aspects are mentioned).

the two best performing extractive (LEXRANK, QT) and abstractive (T5-SIMILAR, ACESUM) systems according to ROUGE. We elicited three judgements for all entities in the SPACE and OPOSUM+ test sets. Table 7 summarizes our results. BWS values range from −100 (unanimously worst) to 100 (unanimously best). ACESUM is deemed best for all criteria on both datasets. Crowdworkers also rated QT high on informativeness, which indicates that aspect modeling is helpful, but low on other criteria (e.g., coherence and conciseness) due to its extractive nature.

Aspect Controllability We also conducted a user study to assess the quality of aspect-specific summaries. We showed participants the aspect in question as well as aspect summaries from T5-SIMILAR, QT, ACESUM, and HUMAN. Crowdworkers were asked to decide whether the summaries discussed the given aspect *exclusively*, *partially*, or *not at all*. We elicited three judgments for all test entities. As can be seen in Table 8, SPACE summaries produced by ACESUM exclusively discuss a single aspect 50.9% of the time. T5-SIMILAR mostly produces general summaries (74.8% of them partially discuss the given aspect) which is not surprising, given that it has no special-purpose mechanism for modeling aspect. QT summaries are more topical for the opposite reason. In general, automatic systems perform worse on OPOSUM+ whose larger number of domains renders this dataset more challenging. Finally, we observe a big gap between model and HUMAN performance.

We further verified whether ACESUM can produce summaries covering two aspects. Although it can generate summaries with more aspects (see Table 1), we hypothesize that user queries pertaining to two aspects would be most frequent. Besides, if performance with two aspects is inferior, there is little chance it will improve with more aspects. For each test example we elicited three judgments and randomly selected two aspect pairs from the set of all possible aspect combinations. We compared ACESUM against QT (for which we used seed words representing both target aspects). Participants were shown the two aspects and the summaries generated by QT and ACESUM. They were asked to decide whether the summaries discussed (a) both target aspects exclusively (b) one of the aspects (c) other aspects in addition to the target ones, and (d) none of the two aspects. The results in Table 9 show that ACESUM is able to produce two-aspect summaries effectively 61.3% of the time on SPACE and 47.0% of the time on OPOSUM+. QT on the other hand mostly creates single-aspect summaries.

Summary Veridicality Our third study examined the veridicality of the generated summaries, i.e., whether the opinions mentioned in them are indeed discussed in the input reviews. Participants were shown reviews and corresponding system summaries and were asked to verify, for each sentence of the summary, whether it was fully supported by the reviews, partially supported, or not at all supported. We performed this experiment

OPOSUM+ General			
Model	FullSupp	PartSupp	NoSupp
T5-SIMILAR	53.3	36.9	9.8
ACESUM	59.9	32.2	8.0
HUMAN	88.4	7.0	4.6

OPOSUM+ Aspect			
Model	FullSupp	PartSupp	NoSupp
T5-SIMILAR	57.3	29.4	13.3
ACESUM	54.2	32.3	13.5
HUMAN	67.8	20.7	11.6

Table 10: *Summary veridicality* evaluation. Proportion of summaries that are fully supported, partially supported, or not supported at all.

on OPOSUM+ only since the number of reviews is small and participants could read them all in a timely fashion. We collected three judgments for all system summaries, both general and aspect-specific ones. Participants assessed the summaries produced by T5-SIMILAR and ACESUM. We also included GOLD-standard summaries as an upper bound but no output from an extractive system as it by default produces veridical summaries which contain facts mentioned in the reviews.

Table 10 reports the percentage of fully (FullSupp), partially (PartSupp), and un-supported (NoSupp) sentences. Not unsurprisingly, GOLD summaries display the highest percentage of fully supported sentences for both general and aspect-specific summaries. ACESUM and T5-SIMILAR present similar proportions of supported sentences when it comes to general summaries, with ACESUM having a slight advantage. The proportion of supported sentences is higher in aspect summaries for T5-SIMILAR. Note that this model struggles to actually generate aspect-specific summaries (see Table 8); instead, it generates any-aspect summaries which maybe veridical but off-topic.

6 Conclusions

In this work, we presented an abstractive approach to aspect-controlled opinion summarization. Key to our model is the induction of aspect controllers which facilitate the creation of a synthetic training dataset and guide summary generation towards the designated aspects. Extensive experiments on two benchmarks show that our model achieves state of the art across the board, for both general and aspect-specific opinion summarization.

In the future, we would like to focus on controlling additional facets of opinion summaries such as sentiment or length. It would also be interesting to learn aspects from data rather than specifying them apriori as well as dealing with unseen aspects (e.g., in a scenario where reviews discuss new features of a product).

Acknowledgments We thank the anonymous reviewers for their feedback. We gratefully acknowledge the support of the European Research Council (Lapata, award number 681760). The first author is supported by a Google PhD Fellowship.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised opinion summarization with content planning. In *Proceedings of the 35th Conference on Artificial Intelligence*, pages 12489–12497. AAAI Press.

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. 2021. Extractive Opinion Summarization in Quantized Transformer Spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Stefanos Angelidis and Mirella Lapata. 2018a. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Stefanos Angelidis and Mirella Lapata. 2018b. 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.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. 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.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. Inference time style control for summarization. In *Proceedings of the 2021 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, pages 5942–5953, Online. Association for Computational Linguistics.

Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. 2013. Mutli-document summarization of evaluative text. *Computational Intelligence*, 29(4):545–576.

Eric Chu and Peter Liu. 2019. MeanSum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232, Long Beach, California, USA. PMLR.

Hoa Trang Dang. 2006. DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 48–55, Sydney, Australia. 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.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71.

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(1):457–479.

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, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. 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.

Sebastian Gehrmann, Falcon Dai, Henry Elder, and Alexander Rush. 2018. End-to-end content and plan selection for data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 46–56, Tilburg University, The Netherlands. Association for Computational Linguistics.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada. Association for Computational Linguistics.

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, New York, NY, USA. Association for Computing Machinery.

Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the web. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1621–1624. AAAI Press.

Jim Keeler and David E. Rumelhart. 1991. A self-organizing integrated segmentation and recognition neural net. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, page 496–503, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. Comprehensive review of opinion summarization. Technical report.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *Proceedings of the 3rd International Conference on Learning Representations*, Banff, AB.

Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597–606, New York, NY, USA.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*, New Orleans, LA, USA.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Oded Maron and Aparna Lakshmi Ratan. 1998. Multiple-instance learning for natural scene classification. In *Proceedings of the Annual International Conference on Machine Learning*, volume 98, pages 341–349.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 43–52, New York, NY, USA. Association for Computing Machinery.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 339–348, Jeju Island, Korea. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. 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 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6309–6318, Red Hook, NY, USA. Curran Associates Inc.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*, Vancouver, Canada.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938.

Colin Raffel, Noam 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. *Journal of Machine Learning Research*, 21(140):1–67.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, USA. Curran Associates Inc.

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, San Diego, California. Association for Computational Linguistics.

Xiu-Shen Wei, Jianxin Wu, and Zhi-Hua Zhou. 2014. Scalable multi-instance learning. In *Proceedings of the 2014 IEEE International Conference on Data Mining*, pages 1037–1042, Shenzhen, China.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yumo Xu and Mirella Lapata. 2019. Weakly supervised domain detection. *Transactions of the Association for Computational Linguistics*, 7:581–596.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. 2009. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1249–1256, Montreal, Quebec, Canada.

Aspect	"Hotels"
building	lobby pool decor gym area
cleanliness	clean spotless garbage dirty stain
food	breakfast food buffet restaurant meal
location	location walk station distance bus
rooms	room bed bathroom shower spacious
service	staff service friendly helpful desk

Table 11: SPACE seed words for the "Hotels" domain.

Aspect	"Laptop Bags"
looks	looks color stylish looked pretty
quality	quality material poor broke durable
size	fit fits size big space

Aspect	"Bluetooth Headsets"
comfort	ear fit comfortable fits buds
ease of use	easy button simple setup control
sound quality	sound quality hear noise volume

Aspect	"Boots"
comfort	comfortable foot hurt ankle comfy
looks	cute look looked fringe style
size	size half big little bigger

Aspect	"Keyboards"
build quality	working months build stopped quality
feel/comfort	feel comfortable feels mushy shallow
layout	key keys delete backspace size

Aspect	"Televisions"
connectivity	hdmi computer port usb internet
image quality	picture color colors bright clear
sound quality	sound speakers loud tinny bass

Aspect	"Vacuums"
accessories	filter brush attachments attachment turbo
ease of use	easy push corners awkward impossible
suction power	suction powerful power hair quiet

Table 12: OPOSUM+ seed words for various domains and their aspects.

A Appendix

A.1 List of Seed Words

Tables 11 and 12 shows the seed words we used in our experiments. These were generated semi-automatically: we first obtained aspect-specific words through the automatic method introduced in Angelidis and Lapata (2018b) and Angelidis et al. (2021) and then asked human annotators to filter out the noise (i.e., words that were assigned incorrect aspects).

A.2 Results using Automatic Seed Words

Table 13 shows comparisons between ACESUM and QT using automatically generated seed words for aspect-specific summarization (as used in Angelidis et al., 2021). Our model performs better than QT on both datasets, while both models benefit from better quality seed words with noticeable increase in ROUGE scores.

	SPACE			OPOSUM+		
Model	R1	R2	RL	R1	R2	RL
using automatic seed words						
QT	28.95	8.34	21.77	23.16	4.13	16.81
ACESUM	30.78	8.39	23.82	27.11	6.05	19.67
using human seed words						
QT	29.43	8.45	22.37	23.99	4.36	16.61
ACESUM	31.80	9.53	25.03	27.55	6.44	20.16

Table 13: ROUGE scores of QT and ACESUM for aspect-specific summarization.

A.3 Extensions to OPOSUM Dataset

In this section, we present our additions to the OPOSUM dataset (Angelidis and Lapata, 2018b). Firstly, we increased the size of the review corpus. The original dataset includes only 359K reviews, which is the result of down-sampling the *Amazon Product Dataset* introduced in McAuley et al. (2015). We instead gathered all reviews tagged with at least one of the OPOSUM domains ("Laptop Bags", "Bluetooth Headsets", "Boots", "Keyboards", "Televisions", and "Vacuums") from the newest version of the *Amazon Product Dataset* compiled by Ni et al. (2019). Since "Laptop Bags" and "Bluetooth Headsets" were significantly smaller than the other four domains, we additionally included all reviews tagged with "Bags" and "Headsets". We were able to increase the dataset to 4.13M reviews, i.e., by a factor of 12.

Secondly, we created a large collection of human-written abstractive summaries for aspect-specific summarization evaluation. For each test product (e.g., television set) and for each aspect (e.g., image quality), we asked three annotators to write an opinion summary about the given aspect. The annotators were shown 10 input reviews, in which opinions about the target aspect were highlighted to aid them in their task. We only used the three most common aspects for each domain, since opinions about less common aspects do not appear frequently in reviews. We gathered 540 aspect-specific summaries in total.

A.4 Example Summaries

Finally, we show general and aspect-specific summaries produced by QT, T5-SIMILAR, ACESUM, and HUMAN on SPACE (Table 14) and OPOSUM+ (Table 15). We also show two-aspect summaries produced by QT and ACESUM in Table 16.

HUMAN summaries	
General	Staff was service focused and very welcoming. Common areas of the hotel smelled fresh because of how clean everything was. The rooms were comfortable and came with a fridge and a microwave. Food, both hot and cold, was very well presented and fresh. The hotel was located within walking distance to the French quarter and felt very safe at night.
Building	It's older, looking at the hotel and lobby, but has lots of charm & character.
Cleanliness	The hotel's lounge, bathrooms, hallways, and even the bedding were all clean and even smelled fresh.
Food	The breakfast is very good and plentiful and was more than just continental, offering eggs, sausage and grits in addition to the usual waffles, cereal, and fruit.
Location	The location is very good, walking distance to all major sights in French quarter.
Rooms	The room is comfortable and equipped with just about everything anyone could need ... a refrigerator, microwave, desk, sofa, iron and ironing board, and hairdryer. The room was also spacious and the hotel was very quiet.
Service	Hotel staff were unbelievably friendly and helpful; they often went above and beyond to be accommodating.

ACESUM summaries	
General	The hotel is in a great location, close to the French quarter and the market. The room was clean and comfortable. Breakfast was good, and the staff was very helpful. There is a small restaurant in the lobby.
Building	The lobby is a bit small. The lobby area is a little bit dated, but the rooms are very comfortable.
Cleanliness	The room was clean and comfortable. The bathroom was very clean with a nice shower.
Food	The breakfast was very good, with a variety of choices. The breakfast buffet was good.
Location	The location is great, right in the heart of Bourbon street, and within walking distance of the French quarter.
Rooms	The room was very spacious and the bathroom was very nice. The room had a TV, a microwave, and a separate shower. There was a small fridge in the room, which was nice.
Service	The staff was very friendly and helpful.

T5-SIMILAR summaries	
General	I stayed at the hotel for 3 nights. The room was very clean, the staff was friendly and the breakfast was excellent! The location was great - a short walk to the Eiffel Tower and St Marks Square.
Building	i stayed here for 3 nights. The staff was very friendly and helpful, the rooms were clean and the location was great! The breakfast was good and there was a lot to do in the city.
Cleanliness	I stayed at the hotel for 3 nights. The staff was very friendly and helpful, the rooms were clean and the breakfast was excellent! The location is great - just a few minutes walk from the Eiffel Tower and it's close to the metro and Bourbon Street.
Food	I stayed at the San Diego for a week. The staff was very friendly and helpful, the rooms were clean and the location was great! The breakfast was good and there wasn't much to do in the room.
Location	I stayed at the hotel for 3 nights. The room was very clean, the staff were very helpful and the breakfast was excellent! The location is great - a short walk to the Eiffel Tower and St Marks Square.
Rooms	I stayed at the hotel for 3 nights. The room was very clean, the staff were very helpful and the location was great! The rooms were clean and well appointed - the breakfast was good and there was a great selection of food and drink options in the morning.
Service	I stayed at the hotel for 3 nights. The room was very clean, the staff was friendly and the breakfast was good! The location was great - a short walk to the Eiffel Tower and St Marks Square.

QT summaries	
General	Great location. The breakfast was very good. We would definitely stay here again. Room was clean. This hotel is great. The room was large with two queen beds. Nice hotel in a nice location. This is a multi-year award winning hotel. Staff were very helpful. The hotel is very clean. Front desk was friendly and helpful. The room was clean and comfy. The breakfast was average. It is very good. We enjoyed our stay here.
Building	Plus all these fancy hotels have the irritating routine of charging around $16 for internet access. The bad: the hotel is quite old and needs renovating.
Cleanliness	Pick this one. Toom was clean. The hotel is very clean. Great 5 star service. Room was nice and clean. This one was by far the best.
Food	The breakfast was very good. When you factor in the delicious complimentary breakfast consisting of scrambled eggs, grits, freshly-made waffles, bagels, bacon, sausage, cereal, toast, juice, and coffee.
Location	But it is just far enough away from the craziness of Bourbon and Canal streets. Walk. The location was also nice. The location of the hotel was excellent.
Rooms	The room was great. The bed was comfortable. The room was large with two queen beds. Hotel room was clean and comfortable.
Service	The staff was very nice. Every member of the staff we encountered was gracious, friendly and helpful. The staff at the hotel are super nice and attentive.

Table 14: *General* and *aspect-specific* summaries for a hotel generated by four systems (SPACE dataset).

HUMAN summaries	
General	The price is great. Lightweight and comfortable fit in the ear. Based on feedback from others I can be heard clearly. Very easy to use and compatible with all of my phones! It holds a charge great, is light enough. The sound quality is great, but cheap.
Comfort	This unit is very lightweight and comfortable to fit in the ear. Makes it really easy to keep in there. Fits perfectly.
Ease of Use	The headphones are very easy to use and fit almost any cell phone. The only problem is that sometimes it is difficult to know if the equipment is on or off.
Sound Quality	The sound from the headphones is very good, the audio quality is excellent. Despite this they could have a louder sound for their maximum volume.

ACESUM summaries	
General	I've had this for a few months now and it's still working great. I haven't had any issues with the sound quality, and I am very happy with the quality of the product. The earpiece is very comfortable and is very lightweight.
Comfort	It's a comfortable fit. The earpiece isn't too big, but it fits your ear well. It is a little snug, but my ear is very comfortable. It is comfortable to wear for long periods of time.
Ease of Use	It is easy to use, easy to set up, and easy to connect to my phone. It's not that easy to get on and off.
Sound Quality	The sound quality is good, but I'm not a fan. I haven't had any issues with the sound of the headphones.

T5-SIMILAR summaries	
General	I bought this headset for my son. He loves it, but a lot of people have complained about the sound quality of the headset! It is very comfortable and the mic isn't loud enough to hear. It has sounded great.
Comfort	I bought this headset for my husband. He loves it, and it works great! It has a great sound and the sound quality is excellent - the only thing is that the microphone isn't very loud.
Ease of Use	I bought this headset for my husband. He loves it, and it is very comfortable! If you are looking for a good headset, this is the best headset you can buy for the price
Sound Quality	I bought these for my husband. He loves them, and they are very comfortable! They don't have a lot of noise. If you are looking for something that will work for you, then they're ok but they will not work with the earbuds.
QT summaries	
General	Thank you! The battery life is ... bizarre. Light to the ear. I highly recommend this bluetooth headset. Lightweight and comfortable fit in the ear. I returned it and received a refund. I used it mostly in my car on my commute to work. Great product.
Comfort	I would really like it if it would stay in my ear or if the loop that went around my ear would hold it to my ear. I could not get this headset to work.
Ease of Use	Item delivery just as described! Its made of the cheapest of materials and the bluetooth has a hard time staying connected. My only gripe is that sometimes there's a small lapse between my voice.
Sound Quality	Also they are comfy and stay in my ears. The headset is light and fits comfortably in my ears (though it takes some time to find the right angle and fit it right in).

Table 15: *General* and *aspect-specific* summaries for the "Bluetooth Headsets" domain) generated by four different systems (OPOSUM+ dataset).

ACESUM summaries
Cleanliness and **Location** of a hotel
The hotel is clean and the rooms are very clean. The location is great, right on the beach, and close to the Eiffel Tower.
Building and **Cleanliness** of a hotel
The room was very clean and the bathroom was very clean. The pool was nice, but the pool area was a bit small.
Food and **Rooms** of a hotel
The breakfast was good, the food was good and the staff was very friendly. The breakfast buffet was good with a variety of choices.
Quality and **Size** of a laptop bag
It's a good size for a laptop. It is not a heavy bag, it is made of a soft material.
Ease of Use and **Suction Power** of a vacuum
I've had this vacuum for a few months now and it's very easy to use. I don't like the fact that it is a little heavy, but it does a great job of picking up the hair.
Comfort and **Looks** of a pair of boots
They are a little tight, and they are not comfortable. They look great with jeans and skirts. If you are looking for a comfortable shoe that will last a long time, do not order this.

QT summaries
Cleanliness and **Location** of a hotel
Overall we had a nice stay at the hotel. It's well worth the extra money. For the price I paid it underwhelmed ($350 for 1 night). Doesn't get more LA than this have a drink at the roof top.
Building and **Cleanliness** of a hotel
The service was great! The hotel was beautiful. amazing. Holy cow. I love staying at this hotel. Excellent. Superb service!! I can't say enough about how perfect this hotel was for us. I stayed at this hotel not too long.
Food and **Rooms** of a hotel
(Note that breakfast isn't necessarily included in the price.) On the first floor there is a small breakfast room but no restaurant. Also a small but cosy terrace with swimming pool. Rooms are a decent size but walls are paper thin.
Quality and **Size** of a laptop bag
The hand straps have not ripped or torn so really I think the problem was that I put too much weight in the bag. Barely fit a 14 inch HP sleek notebook. I would not recommend this bag
Ease of Use and **Suction Power** of a vacuum
I even tried putting ear plugs in to vacuum with it, but it still hurts my ears. I looked at every small but powerful vacuum I could find in stores and on line.
Comfort and **Looks** of a pair of boots
Once the weather got cold the shoes became more stiff and they really hurt now so it looks like I wasted $40. I am wondering if they are worth returning or just passing off to someone

Table 16: Opinion summaries focusing on *two aspects* (SPACE and OPOSUM+ datasets)