Simple Yet Effective Synthetic Dataset Construction for Unsupervised Opinion Summarization

Ming Shen[♥*] **Jie Ma**[♣] **Shuai Wang**[♣] **Yogarshi Vyas**[♣]
Kalpit Dixit[♣] **Miguel Ballesteros**[♣] **Yassine Benajiba**[♣]
[♥]Arizona State University [♣]AWS AI Labs
mshen16@asu.edu;{jieman,wshui,yogarshi,kddixit,ballemig,benajiy}@amazon.com

Abstract

Opinion summarization provides an important solution for summarizing opinions expressed among a large number of reviews. However, generating aspect-specific and general summaries is challenging due to the lack of annotated data. In this work, we propose two simple yet effective unsupervised approaches to generate both aspect-specific and general opinion summaries by training on synthetic datasets constructed with aspect-related review contents. Our first approach, *Seed Words Based Leave-One-Out* (SW-LOO), identifies aspect-related portions of reviews simply by exact-matching aspect seed words and outperforms existing methods by 3.4 ROUGE-L points on SPACE and 0.5 ROUGE-1 point on OPOSUM+ for aspect-specific opinion summarization. Our second approach, *Natural Language Inference Based Leave-One-Out* (NLI-LOO) identifies aspect-related sentences utilizing an NLI model in a more general setting without using seed words and outperforms existing approaches by 1.2 ROUGE-L points on SPACE for aspect-specific opinion summarization and remains competitive on other metrics.

1 Introduction

Customer reviews play a vital role in decision-making for customers and product (or business) providers, as customers usually resort to reviews to guide their purchasing decisions and product providers improve their products based on reviews as feedback. However, it becomes hard for customers or product providers to read through all reviews before making decisions with the explosion of online reviews in recent years. Opinion summarization (Hu and Liu, 2006; Wang and Ling, 2016; Angelidis and Lapata, 2018; Bražinskas et al., 2020; Brazinskas et al., 2022; Angelidis et al., 2021; Amplayo et al., 2021a; Basu Roy Chowdhury et al., 2022), the task of generating a general summary of *salient* opinions expressed among reviews, provides a feasible solution to this problem.

Different from summarization in Wikipedia and news domains (Nallapati et al., 2016; Narayan et al., 2018a; See et al., 2017; Narayan et al., 2018b; Liu and Lapata, 2019; Cachola et al., 2020), opinion summarization cannot rely on reference summaries for model training since it is difficult and expensive to annotate large scale reviews-summary pairs. Also, customers usually care about specific aspects of a product instead of a *general* high-level summary. Thus, fine-grained *aspect-specific* opinion summaries are required, and this makes the annotation process even more difficult and expensive.

Amplayo et al. (2021a) propose an abstractive approach to generate aspect-specific opinion summaries by training on synthetic datasets. They construct synthetic datasets with review elements (words, phrases, or sentences) identified by a multiple instance learning (MIL) module (Keeler and Rumelhart, 1991) learned with silver-standard labels obtained using aspect seed words. We first follow this direction to propose a more straightforward and effective method that excludes the complex learning module to identify aspect-related elements to construct synthetic datasets. Moreover, aspect seed words, which again require human efforts, may not always be available when moving to new domains. Thus we propose another more general solution without the curation and supervision of aspect seed words.

Specifically, we propose two simple yet effective methods to identify aspect-related review sentences and construct aspect-specific synthetic datasets in a *Leave-One-Out* (LOO) (Bražinskas et al., 2020; Elsahar et al., 2021; Brazinskas et al., 2022) style and then finetune pretrained language models (PLMs) on the synthetic datasets: (a) SW-LOO identifies aspect-related sentences by simply exact-matching aspect seed words and outperforms existing ap-

[*]Work done during an internship at AWS AI Labs.

Findings of the Association for Computational Linguistics: EACL 2023, pages 1898–1911

proaches by 3.4 ROUGE-L points and 0.5 ROUGE-1 point on aspect opinion summaries of SPACE and OPOSUM+ respectively; (b) NLI-LOO identifies aspect-related sentences with a finetuned NLI (Bowman et al., 2015; Williams et al., 2018) model. Being the first approach that does not use aspect seed words, it outperforms existing approaches on aspect opinion summarization by 1.2 ROUGE-L points for SPACE and falls behind at most 1 ROUGE point on other metrics.

2 Problem Formulation

Let C denote a corpus of reviews on entities $\{e_1, e_2, \dots\}$ (products or business). Let $A_e = \{a_1, a_2, \dots, a_M\}$ denotes a set of aspects (e.g., *food* or *location* for a hotel) that are relevant for the domain of entities. For each entity e, we define its review set as $R_e = \{r_1, r_2, \dots, r_N\}$. Each review r is a collection of sentences $\{x_1, x_2, \dots\}$ and each sentence x is a sequence of tokens $\{w_1, w_2, \dots\}$. Each aspect a is represented by a small set of *seed words* (e.g., *meal* or *buffet* for *food* aspect) $S_a = \{v_1, v_2, \dots\}$. Our approaches generate two types of opinion summaries: (a) *general* summary that contains salient opinions over *all* aspects of the entities; and (b) *aspect* summary that focuses on only one specific aspect $a \in A_e$.

3 Synthetic Dataset Construction

Leave-One-Out (LOO) We construct synthetic datasets in a LOO style: from a pool of review elements (reviews or review sentences), an element is randomly sampled as a *pseudo-summary*, then we select input reviews from the remaining review elements.

3.1 Seed Words Based LOO

To build a synthetic reviews-summary pair for aspect a, as shown in the upper diagram of Figure 1, we first filter each review r into its aspect-related portion r' where $r' \subseteq \{x_1, x_2, \dots\}$ with each sentence in r' containing at least one seed word in A_e. For example, for *food* aspect with its seed words {*breakfast*, *buffet*, ...}, a hotel review r'_i: *"They have the most wonderful buffet in Bay Area. And the hotel is close to the airport. Forgot to mention, especially the breakfast is terrific."* will be filtered into its aspect-related review portion r'_i: *"They have the most wonderful buffet in Bay Area. Forgot to mention, especially the breakfast is terrific."*. Noticed that r'_2 is empty suppose there is no sentence in r_2

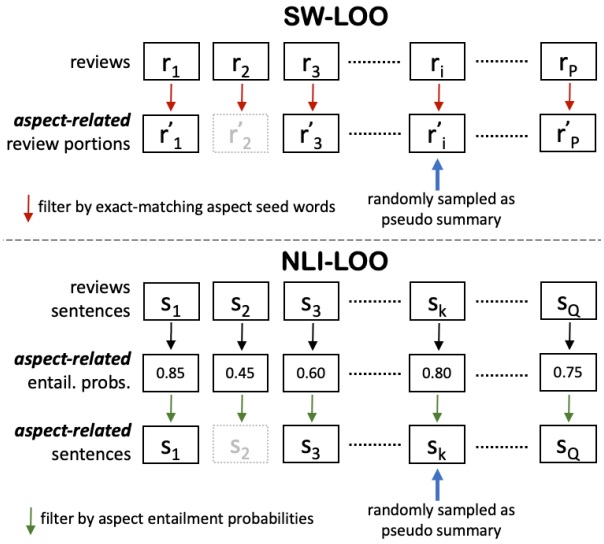

Figure 1. One synthetic data pair construction for aspect a in SW-LOO and NLI-LOO.

containing any seed word. Then we apply LOO construction on the filtered aspect-related review portions $\{r'_1, r'_3, \dots, r'_P\}$ as shown in the diagram: r'_i is randomly sampled as the pseudo summary and inputs are chosen from $\{r'_1, r'_3, \dots, r'_P\} \setminus \{r'_i\}$ by first ranking them with the pseudo-summary r'_i based on ROUGE-1 score (Lin, 2004) and then truncating with a token budget j (truncate up to j tokens) since a concatenation of all filtered reviews cannot fit into the encoder of a PLM. Please refer to Appendix B for more details and analysis on SW-LOO.

3.2 NLI Based LOO

NLI Component In order to relax the requirement of aspect seeds (provided by humans) and to make a more scalable and general solution, we propose to use an NLI model to infer whether a review sentence is related to an aspect. Specifically, we set a review sentence as the premise and verbalize an aspect with the template: *the text is about {aspect}*, which we use as the hypothesis. If the entailment probability is higher than a threshold (0.9 for SPACE and 0.8 for OPOSUM+), we identify the sentence as related to the aspect with this entailment probability, else we set the aspect-related probability to 0.

To build a synthetic pair for aspect a, we first break all reviews into review sentences and filter out those that are not related to aspect a with the NLI model. As shown in the lower diagram of Figure 1, each sentence is first passed through the NLI model to infer its probability of relatedness

to aspect a, so s_2 with entailment probability of 0.45 will be filtered out if the threshold is set to 0.5. Then we apply LOO construction on all aspect-related sentences $\{s_1, s_3, \ldots, s_Q\}$ and we also use a token budget to truncate ranked synthetic input similar to SW-LOO, however, different from SW-LOO where we use ROUGE-1 scores to rank, we calculate similarities based on entailment probabilities. Please refer to Appendix C for more details and analysis on NLI-LOO. Note that we filter the input reviews at sentence level for NLI-LOO and at review level in SW-LOO.

4 Summarization Model

We use T5 (Raffel et al., 2020), a sequence-to-sequence Transformer-based (Vaswani et al., 2017) PLM, to finetune our synthetic datasets similar to previous works (Ke et al., 2022; Amplayo et al., 2021a). For SW-LOO, we use the following template: "`summarize based on aspect: [ASPECT]` *{aspect}* `[ASPECT] with seed words: [SEED]` *{seed words}* `[SEED]:` *{filtered review}* `[SEP]` *{filtered review}* ... " to convert synthetic input and for NLI-LOO, we use: "`[ASPECT]` *{aspect}* `[SEP]` *{aspect-related sent}* `[SEP]` *{aspect-related sent}* ... ". `[ASPECT]`, `[SEED]`, and `[SEP]` are special tokens, *{aspect}* is an aspect name, *{seed words}* are concatenation of seed words for an aspect, each *{filtered review}* is a r_i' in SW-LOO synthetic input, and each *{aspect-related sent}* is a s_k in NLI-LOO synthetic input. For both methods, outputs are pseudo summaries.

5 Experiment

5.1 Datasets

We evaluate our methods on two opinion summarization datasets: SPACE (Angelidis et al., 2021), containing reviews from *hotel* domain, and OPOSUM+ (Amplayo et al., 2021a), containing Amazon product reviews from six different domains. Both datasets are comprised of a large corpus of raw reviews and a small development and test set with human-annotated aspects and general opinion summaries for evaluation. Aspect seed words are usually obtained with a small amount of human effort. For SW-LOO, we use the same seed words as in Amplayo et al. (2021a) (Appendix E). Refer to Appendix D for detailed descriptions and statistics of the two datasets.

	Model	SPACE R1	SPACE R2	SPACE RL	OPOSUM+ R1	OPOSUM+ R2	OPOSUM+ RL
Extractive	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
	ACESUM$_{\text{EXT}}$	30.91	8.77	23.61	26.16	5.75	18.55
	SEMAE	31.24	10.43	24.14	-	-	-
	SW-LOO$_{\text{EXT}}$	<u>33.14</u>	10.32	25.81	28.14	6.10	19.51
	NLI-LOO$_{\text{EXT}}$	27.18	6.63	20.60	26.78	6.48	18.07
Abstractive	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
	ACESUM	32.41	9.47	25.46	<u>29.53</u>	<u>6.79</u>	**21.06**
	SW-LOO	**34.68**	**11.50**	**28.83**	**30.00**	**6.92**	<u>20.76</u>
	NLI-LOO	31.57	<u>10.44</u>	<u>26.66</u>	28.90	6.60	20.11
	HUMAN	44.86	18.45	34.58	43.03	16.16	31.53

Table 1. Evaluation for *aspect summaries* on SPACE and OPOSUM+ test sets. Best performances are in **bold** and second best performances are <u>underlined</u>.

5.2 Baselines

We compare our methods with several unsupervised extractive and abstractive approaches. Extractive approaches include CENTROID (Radev et al., 2004), LEXRANK (Erkan and Radev, 2004), QT (Angelidis et al., 2021), SEMAE (Basu Roy Chowdhury et al., 2022), and two extractive variants of our methods, SW-LOO$_{\text{EXT}}$ and NLI-LOO$_{\text{EXT}}$, by feeding identified aspect-related sentences to LEXRANK instead of T5, similar to the idea in Amplayo et al. (2021a). Abstractive approaches include MEANSUM (Chu and Liu, 2019), COPYCAT (Bražinskas et al., 2020), and ACESUM (Amplayo et al., 2021a). Appendix F contains more details on baselines.

We also compare with two upper bounds reported in Amplayo et al. (2021a): an ORACLE that selects the review with the highest ROUGE score to the gold summary as the summary and a HUMAN upper bound that is calculated as the inter-annotator ROUGE scores.

5.3 Implementation

We first pre-process the raw corpus such as removing products with very few reviews and too long or short reviews as in Appendix G. We use T5-SMALL as our summarization models and larger T5 size does not show improvements as shown in Appendix L. We use a MNLI (Williams et al., 2018) finetuned BART-LARGE (Lewis et al., 2020) model in NLI-LOO. We choose this model given its better performance[1] in zero-shot topic classification. We perform simple hyper-parameter tuning

[1] https://joeddav.github.io/blog/2020/05/29/ZSL.html

	Model	SPACE R1	SPACE R2	SPACE RL	OPOSUM+ R1	OPOSUM+ R2	OPOSUM+ RL
Extractive	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
	ACESUM$_{\text{EXT}}$	35.50	7.82	20.09	38.48	15.17	22.82
	SEMAE	**43.46**	**13.48**	**26.40**	-	-	-
	SW-LOO$_{\text{EXT}}$	38.44	11.01	25.62	**40.45**	**19.13**	23.20
	NLI-LOO$_{\text{EXT}}$	25.07	4.52	16.16	39.79	18.33	**23.49**
Abstractive	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
	ACESUM	40.37	11.51	23.23	32.98	10.72	20.27
	SW-LOO	42.27	12.99	23.47	36.19	12.17	21.11
	NLI-LOO	41.25	12.79	24.31	31.22	9.93	19.08
	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 2. Evaluation for *general summaries* on SPACE and OPOSUM+ test sets. Best performances are highlighted in bold and second-best performances are underlined.

Model	SPACE Aspect	SPACE General	OPOSUM+ Aspect	OPOSUM+ General
SW-LOO	27.59	23.42	20.41	20.58
w/ Training Random	24.24	24.70	19.75	18.71
w/ Inference Random	23.46	22.04	18.76	19.41
w/ Both Random	14.71	21.82	18.06	18.15
NLI-LOO	25.92	25.13	19.21	19.32
w/ Training Random	22.05	22.06	18.42	19.37
w/ Inference Random	24.33	24.56	18.10	16.97
w/ Both Random	16.14	22.50	17.83	19.69

Table 3. *Training Random* means randomly selecting sentences as pseudo summary and input during synthetic dataset construction. *Inference Random* means randomly selecting sentences as input during inference. We report RL scores of our approaches on dev sets.

on dev sets and select checkpoints with the best ROUGE-L scores to report performances on test sets. Please refer to Appendix H for more details such as training configurations and other analyses.

5.4 Results

We evaluate the quality of generated opinion summaries using ROUGE1/2/L F1 scores (Lin, 2004). Example summaries generated by our methods are shown in Table 12 and Table 13 in Appendix.

Aspect Opinion Summarization Table 1 contains the results of all baselines and our methods on the two benchmark datasets. Despite its simplicity, SW-LOO achieves the highest scores on both datasets across all metrics except RL for OPOSUM+ with only 0.3 points behind the best-performing baseline. On the other hand, NLI-LOO achieves higher R2 and RL scores on SPACE than existing methods despite using no seed words. While it falls behind other methods on OPOSUM+, it is at most 1 point behind across all metrics. This highlights that even without aspect seed words, NLI-LOO is possible to compete with SOTA aspect-based opinion summarization methods.

Next, we turn to the evaluation of extractive versions of our methods. We observe SW-LOO$_{\text{EXT}}$ achieves higher R1 and RL scores on SPACE but falls behind on OPOSUM+ by at most 1.5 point compared with all baselines. This is consistent with the finding in Amplayo et al. (2021a) that a simple centrality-based extractive approach such as LEXRANK are strong baselines as long as input sentences are already aspect-related. And SW-LOO$_{\text{EXT}}$ outperforming ACESUM$_{\text{EXT}}$ further shows that our simple filtering method using exact-matching seed words already produces good enough aspect-related sentences compared with the extra learning module used in Amplayo et al. (2021a). However, NLI-LOO$_{\text{EXT}}$, is not able to outperform the best baseline, and we hypothesize the reason is that NLI model filtered aspect-related sentences are still noisy so that a summarization model is required to serve as regularization.

Finally, comparing our four methods, SW-LOO achieves the best performances with the supervision of seed words, NLI-LOO comes second despite the lack of seed words supervision, and our two extractive versions come last since the ground truth summaries are in nature abstractive.

General Opinion Summarization As shown in Table 2, on SPACE, SW-LOO and NLI-LOO outperform the SOTA abstractive system, ACESUM, but under-perform SOTA extractive system, SEMAE. We observe the same trend between SW-LOO$_{\text{EXT}}$ and ACESUM$_{\text{EXT}}$ as in aspect opinion summarization and this again shows the simple yet effective nature of our filtering method. For OPOSUM+, SW-LOO$_{\text{EXT}}$ and NLI-LOO$_{\text{EXT}}$ outperform existing methods given that the annotated general summaries for OPOSUM+ are extractive, SW-LOO outperforms existing abstractive approaches, and NLI-LOO falls behind with only 1 point.

5.5 Ablation Study

We conduct ablation experiments with random filtering to study the importance of the filtering strategies in our two methods. We introduce randomness in two different phases. First, when constructing synthetic pairs, instead of using our filtering strategies before applying LOO construction, we ran-

domly select sentences as pseudo-summary and input. This is essentially a random LOO baseline. Second, during inference, we sample random sentences to feed into T5 encoder instead of using our filtering strategies to select aspect-related elements. Finally, we combine these two random strategies. Results in Table 3 show that our sentence filtering strategies are crucial since ROUGE scores drop drastically as more randomness is introduced. This is more severe for aspect summarization since aspect-specific synthetic dataset construction needs to focus on particular aspects. However, randomly selecting sentences is possible to cover most aspects by chance for general summarization.

6 Conclusion

In this work, we propose two simple yet effective unsupervised approaches that generate aspect and general opinion summaries by training on synthetic datasets. SW-LOO constructs synthetic datasets simply by exact-matching aspect seed words and outperforms existing methods consistently on all metrics and datasets. Being the first work that generates aspect summaries without using aspect seed words, NLI-LOO constructs synthetic datasets with an out-of-the-box NLI model and achieves on-par and sometimes even better performances compared with existing methods.

Limitations

One of the biggest challenge in opinion summarization is the multi-document setting where each document represents one product review. Since the number of reviews for a product tends to be large, it would be unrealistic to concatenate all input reviews and train to generate a summary in an end-to-end fashion limited by modern hardware capacity, for example, the GPU memory needed is quadratic w.r.t the input length for all transformer-based PLM. In this work, we tackle this problem by pre-filtering reviews using some heuristics (aspect seed words matching and NLI model selecting) into sub-elements of reviews with much smaller sizes. However, information is very likely to get lost and become incomplete in the pre-filtering phase, leading to inaccurate summarization. Our approach is exactly facing this problem. One way to address this drawback is to first condense each review into an encoding that contains key information of the review such as opinion aspect and opinion sentiment, and then aggregate all review vectors to generate a summary. Amplayo and Lapata (2021) call this pipeline as CONDENSE-ABSTRACT and it has been used in both supervised and unsupervised general opinion summarization (Chu and Liu, 2019; Coavoux et al., 2019; Iso et al., 2021; Amplayo and Lapata, 2021; Isonuma et al., 2021).

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021a. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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.

Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and controllable opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2662–2672, 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. 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.

Somnath Basu Roy Chowdhury, Chao Zhao, and Snigdha Chaturvedi. 2022. Unsupervised extractive opinion summarization using sparse coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1209–1225, Dublin, Ireland. Association for Computational Linguistics.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*,

pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. 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.

Arthur Brazinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. 2022. Efficient few-shot fine-tuning for opinion summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1509–1523, Seattle, United States. 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.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI*.

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

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *SIGIR '02*.

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.

Bogdan Dumitrescu and Paul Irofti. 2018. *Dictionary learning algorithms and applications*. Springer.

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. *J. Artif. Intell. Res.*, 22:457–479.

Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the web. In *Aaai*, volume 7, pages 1621–1624.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex Aggregation for Opinion Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3885–3903, Punta Cana, Dominican Republic. 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.

Wenjun Ke, Jinhua Gao, Huawei Shen, and Xueqi Cheng. 2022. Consistsum: Unsupervised opinion summarization with the consistency of aspect, sentiment and semantic. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 467–475.

Jim Keeler and David Rumelhart. 1991. A self-organizing integrated segmentation and recognition neural net. *Advances in neural information processing systems*, 4.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

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.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

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*, SIGIR '15, page 43–52, New York, NY, USA. Association for Computing Machinery.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. 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.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking : Bringing order to the web. In *WWW 1999*.

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

Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manag.*, 40:919–938.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sascha Rothe, Joshua Maynez, and Shashi Narayan. 2021. A thorough evaluation of task-specific pretraining for summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 140–145, Online and Punta Cana, Dominican Republic. 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.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. 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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Related Works

Unsupervised opinion summarization is the task of summarizing opinionated text such as customer reviews without training on gold reviews-summary pairs. Recent works have been using autoencoders (Kingma and Welling, 2014) and synthetic datasets construction, or a mix of both, to tackle the zero-shot setting.

An autoencoder model consists of an encoder that maps the input into latent embedding space and a decoder that reconstructs the original input from the latent space. The latent representation learned can be later aggregated or can be used to cluster and select text to perform both extractive and abstractive summarization. Chu and Liu (2019); Bražinskas et al. (2020) aggregate the input reviews latent representations by averaging then generate the summaries conditioned on it. Angelidis and Lapata (2018) utilizes the latent representation with aspect specificity and sentiment polarity to guide the selection of review texts as extractive summaries. Recently, Angelidis et al. (2021) proposes the first approach that generates both general and *aspect-specific* opinion summaries in an extractive manner. They first leverage Vector-Quantized Variational Autoencoder (Van Den Oord et al., 2017) to cluster review sentences and then use a popularity-driven extraction algorithm to summarize. Similar to Angelidis et al. (2021), Basu Roy Chowdhury et al. (2022) learns representations of texts over latent semantic units using dictionary learning (Dumitrescu and Irofti, 2018). Other autoencoder-related methods include denoising autoencoder (Amplayo and Lapata, 2020) and Coavoux et al. (2019), an encoder-decoder architecture that utilizes clustering of encoding space to extract summaries.

Another direction of work creates synthetic datasets utilizing the largely available amount of online customer reviews. Synthetic datasets are usually constructed in a *leave-one-out* (LOO) style that one review is first randomly sampled as a pseudo-summary, and then a subset of reviews are selected or generated as input reviews to be paired with the pseudo-summary to enable supervised training. Methods of selecting and generating input reviews include random sampling (Bražinskas et al., 2020), generating noisy versions of the pseudo-summary (Amplayo and Lapata, 2020), selecting reviews that have closer distribution with the pseudo-summary in the embedding space (Amplayo et al., 2021b; Ke et al., 2022), and selecting more textual similar reviews (Elsahar et al., 2021; Brazinskas et al., 2022). Recently, Amplayo et al. (2021a) proposes the first abstractive approach that can generate both general and aspect summaries. Their method build synthetic datasets by identifying aspect-specific elements with a multiple instance learning (MIL) model (Keeler and Rumelhart, 1991) using aspect seed words. Our work is closest to Amplayo et al. (2021a) in that we also build synthetic datasets by identifying aspect-specific elements, however, our methods do not require extra learning components such as MIL but achieve better performances.

Besides unsupervised opinion summarization, our second method, NLI-LOO is related to the recent approach (Yin et al., 2019) that utilizes NLI (Bowman et al., 2015; Williams et al., 2018) models to tackle zero-shot text classification (Chang et al., 2008) (multi-class and multi-label) problem such as topic detection (Zhang et al., 2015) and emotion detection (Bostan and Klinger, 2018). The main idea is to solve the classification problem by casting the problem into NLI format. Specifically, the text to be classified becomes the premise, and class labels are converted into natural language format (verbalization) to be used as the hypothesis. If the text entails the verbalized class label, then the text belongs to this class. In our work, we identify the relatedness of a review sentence to an aspect in such a way to construct synthetic datasets.

B SW-LOO Details

For general synthetic pairs construction, after filtering each review with seed words for each aspect, we make sure to sample one review such that its aspect-related portions for all aspects are non-empty and concatenate them as pseudo-summary. We retrieve top similar filtered reviews to each aspect-related portion in the pseudo-summary and concatenate them as general synthetic input, and the retrieval process is the same as in aspect synthetic pairs construction. General synthetic input and output are both approximately M times the length of those in aspect synthetic pairs where M is the number of aspects. For the summarization model, *{aspect}* and *{seed words}* are the concatenation of all aspects and all seed words for general synthetic pairs. Finally, we train all synthetic pairs together.

At inference time, we also first filter each review

into aspect-related portions. However, since there is no reference pseudo summary, we cannot truncate based on similarities to fit into T5 encoder. We adopt the *principle strategy* used in PEGASUS (Zhang et al., 2020) Gap Sentences Generation pre-training objective to select important reviews as input for inference. We show the effectiveness of adopting the principle strategy in Appendix I.

C NLI-LOO Details

Different from SW-LOO where we use ROUGE-1 scores, we calculate similarities based on aspect entailment probabilities to rank and truncate aspect-related sentences as synthetic input. For aspect synthetic pairs, we simply calculate the absolute probability difference between pseudo summary and aspect-related sentences. For general synthetic pairs, each review sentence (no matter whether aspect-related) corresponds to a probability vector of dimension M where M is the number of aspects and each element is the probability of the sentence being related to each aspect, and we calculate cosine similarities between the probability vectors of pseudo summary and review sentences that are related to at least one aspect (sum of the probability vector is non-zero). We use the same token budget to truncate review sentences to fit into T5 encoder for both aspect and general synthetic pairs. We also train all synthetic pairs together. Another way to calculate similarities is directly using cosine similarity between sentence embeddings, however, results reported in Appendix J do not show better performance.

During inference, we use 1 and all-one vectors with dimension M as reference vectors to rank and truncate review sentences for aspect and general input construction.

D Datasets Details

Hotel reviews in SPACE are collected from TripAdvisor and each hotel in the evaluation sets is annotated with seven types of summaries: six aspect-specific and one general, with three gold summaries for each type. The number of reviews for a hotel in the raw corpus varies but each hotel in the evaluation sets comes with 100 reviews. Product reviews from six domains: *laptop bag*, *Bluetooth headset*, *boots*, *keyboard*, *television*, and *vacuum* in OPOSUM+ are initially down-sampled from *Amazon Product Dataset*[2] (McAuley et al., 2015) by Angelidis and Lapata (2018) and then further expanded by Amplayo et al. (2021a). Each product in the evaluation sets is annotated with four types of summaries: three aspect-specific and one general, with also three gold summaries for each type. The number of reviews for a product in the raw corpus also varies but each product in the evaluation sets comes with 10 reviews. All human-annotated summaries are abstractive except that general summaries in OPOSUM+ are extractive. Detailed statistics of the datasets are shown in Table 4.

[2] http://jmcauley.ucsd.edu/data/amazon/

Statistics	SPACE	OPOSUM+
domain	1	6
aspects per entity	6	3
raw review corpus		
entities	11.40K	95.55K
total reviews	1.14M	4.13M
dev / test set		
entities	25	30
reviews per entity	100	10
summaries per entity	3	3
total aspect summaries	450	270
total general summaries	75	90

Table 4. Detailed statistic for SPACE and OPOSUM+ datasets. Note that only gold general summaries for OPOSUM+, which is underlined in the table, are extractive.

E List of Seed Words

Aspect seed words (listed in Table 5 and 6) are usually automatically extracted using a variant of clarity scoring function (Cronen-Townsend et al., 2002) applied on a small amount of aspect annotation as described in Angelidis and Lapata (2018), and they can be further manually improved by domain experts as in Amplayo et al. (2021a).

F Baselines Details

Extractive Approaches We first compare against two traditional approaches: CENTROID selects the review closest to the centroid of all reviews as the summary; LEXRANK selects the most salient review sentences as summary similar to PAGERANK (Page et al., 1999). BERT (Devlin et al., 2019) embedding is used to represent sentences in both traditional methods. More recent systems include QT (described in Section 1) and SEMAE. Inspired by QT, SEMAE represents text over latent semantic units using dictionary learning.

Abstractive Approaches MEANSUM generates summaries by reconstructing the mean of reviews'

Aspect	Hotel
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 5. Seed words for *hotel* domain in SPACE dataset.

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

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

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

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

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

Aspect	Vacuum
accessory	filter brush attachments attachment turbo
ease of use	easy push concerns awkward impossible
suction	suction powerful power hair quiet

Table 6. Seed words for various domains in OPOSUM+ dataset.

representations using autoencoder. COPYCAT uses a hierarchical variational autoencoder to learn latent codes for the summaries. The most recent approach is ACESUM. (described in Section 1).

Note that LEXRANK, MEANSUM, and COPYCAT do not support aspect-specific summary generation, Amplayo et al. (2021a) adopt a simple sentence-filtering strategy to enable it. Specifically, after training a general opinion summarization model, during inference for aspect summaries, they filter out input review sentences that are not aspect-related using cosine similarities scores between BERT embeddings of review sentences and aspect seed words before feeding into general summarization model.

G Datasets Pre-Processing

We pre-process differently for our two methods on the same dataset since we want to control the constructed synthetic datasets to have reasonable sizes and resemble properties of test time data such as the number of reviews per product and average review length. We use dev sets to observe such properties. In SW-LOO, we first remove reviews with less than 20 words and then remove hotels with less than 10 reviews for SPACE; we first remove reviews less than 20 or more than 100 words then remove products with less than 12 reviews for OPOSUM+. In NLI-LOO, we remove reviews with less than 10 or more than 120 words for SPACE and remove reviews with less than 20 or more than 100 words for OPOSUM+.

H Implementation Details

We use T5 implementation from HuggingFace[3] (Wolf et al., 2020). We use `AdamW` (Loshchilov and Hutter, 2019) optimizer without weight decay and set 0.9, 0.999, 1×10^{-8} for β_1, β_2, ϵ. We train all summarization models for a total of 25K steps on the combination of aspect and general synthetic pairs. We set ngram refraining size (Paulus et al., 2018) to 3 during inference. We tune initial learning rate in [1e−6, 4e−5, 3e−4] and batch size in [8, 16]. We tune beam search size during inference in [2, 4]. For SW-LOO$_{\text{EXT}}$ and NLI-LOO$_{\text{EXT}}$, we use `[CLS]` token embedding in the last layer of BERT as the sentence representation. We concatenate top 6 sentences returned by LEXRANK as general summary, and tune in [2, 4] for aspect summary. We also use two sizes of BERT: BERT-BASE and BERT-LARGE. All computations are performed on 8-GPU p3.16xlarge Amazon instance. The best hyper-parameter settings for all experiments can be found in Table 7.

During preliminary studies for aspect synthetic pairs construction, we find that for SPACE, us-

[3] https://huggingface.co/docs/transformers/model_doc/t5

SW-LOO		
SPACE	asp.	lr=3e−4, bch=16, bm=2
	gen.	lr=3e−4, bch=16, bm=2
OPOSUM+	asp.	lr=3e−4, bch=16, bm=2
	gen.	lr=1e−6, bch=16, bm=2
NLI-LOO		
SPACE	asp.	lr=4e−5, bch=16, bm=2
	gen.	lr=4e−5, bch=16, bm=2
OPOSUM+	asp.	lr=3e−4, bch=8, bm=4
	gen.	lr=1e−6, bch=16, bm=2
SW-LOO$_{\text{EXT}}$		
SPACE	asp.	BERT-Base, n=2
	gen.	BERT-Base, n=2
OPOSUM+	asp.	BERT-Base, n=2
	gen.	BERT-Large, n=2
NLI-LOO$_{\text{EXT}}$		
SPACE	asp.	BERT-Large, n=2
	gen.	BERT-Base, n=4
OPOSUM+	asp.	BERT-Large, n=4
	gen.	BERT-Large, n=4

Table 7. Best hyper-parameter settings on SPACE and OPOSUM+ dev sets: lr stands for AdamW initial learning rate, bch stands for training batch size, and bm stands for beam search size at inference time.

ing sampled filtered aspect-related review portion as pseudo-summary rather than the original review that contains the pseudo-summary gives better downstream ROUGE scores, but it is the opposite way with OPOSUM+. Please refer to Appendix K for analyses on pseudo-summary granularity.

SW-LOO For SPACE, we add a linear learning rate warm-up in the first 500 steps and save checkpoints every 500 steps. Since there are totally 6 aspects for SPACE and very few reviews containing seed words for all 6 aspects can be sampled as pseudo summaries for general synthetic pairs construction, we relax the constraint of pseudo summaries containing seed words for all 6 aspects to 4 aspects. We set 200 as the token budget to truncate ranked aspect-related review portions for aspect synthetic pairs construction, and 150 as the token budget in principle strategy when selecting important sentences as input during inference for SPACE. We set 1536 and 200 as the maximum input and output token length of T5 for all SW-LOO experiments. Notice that this exceeds 512, which is the maximum token length that T5 is pretrained on, but recent works (Zhang et al., 2020; Rothe et al., 2021) have shown that seq2seq PLMs generalize well even when finetuned on longer sequences not observed at pretraining phase. For OPOSUM+, we add linear learning rate warm-up in the first 250 steps. We set 300 as the token budget to truncate for aspect synthetic pairs construction. There are ~ 50K aspect and ~ 5K general synthetic pairs for SPACE, ~ 70K aspect and ~ 6K general synthetic pairs for OPOSUM+.

Model	Aspect R1	Aspect R2	Aspect RL	General R1	General R2	General RL
SW-LOO	33.11	10.98	27.59	40.26	12.04	23.42
w/o Prin. Sel.	30.88	9.74	25.78	35.84	10.28	21.80

Table 8. Randomly or using principle strategy to select aspect-related review portions in order to fit into the encoder of T5. Performances are reported on SPACE dev set.

NLI-LOO For SPACE, we add linear learning rate warm-up to first 1000 steps, and 500 steps for OPOSUM+. We set 0.9 and 0.8 as the entailment probability threshold for SPACE and OPOSUM+ based on our preliminary experiments (lower thresholds make identified aspect-related sentences too noisy and further hurt downstream ROUGE scores). For summarization models, we set 500 as the token budget for both aspect and general synthetic pairs construction for both datasets and set 512 and 150 for maximum input and output token length of T5. There are ~ 36K aspect and ~ 6K general synthetic pairs for SPACE, ~ 70K aspect-specific and ~ 28K general synthetic pairs for OPOSUM+.

I Principle Strategy Effectiveness

Unlike OPOSUM+, there are 100 reviews for each hotel in SPACE evaluation sets. During inference, we cannot simply concatenate all filtered reviews as input since they cannot fit into T5 encoder. We adopt the principle strategy introduced in PEGASUS to select the most important filtered reviews and concatenate them as input for inference. In Table 8, we show the effectiveness of the principle strategy by comparing it with randomly selecting filtered reviews as input for inference.

J Similarity Metric

Different from using aspect entailment probability, we can also use sentence embeddings (Reimers and Gurevych, 2019) to calculate the cosine similarity between pseudo summary and aspect-related review sentences to construct synthetic input. Specifi-

	Model	SPACE R1	SPACE R2	SPACE RL	OPOSUM+ R1	OPOSUM+ R2	OPOSUM+ RL
Asp.	NLI-LOO	30.20	9.84	25.92	27.48	5.64	19.21
	w/ Sent. Sim.	29.87	9.30	25.26	27.00	6.20	19.06
Gen.	NLI-LOO	41.17	12.34	25.13	31.10	10.09	19.32
	w/ Sent. Sim.	25.01	9.68	17.34	31.11	10.43	19.86

Table 9. Calculate cosine similarity using aspect entailment probability or sentence embeddings when constructing synthetic datasets in NLI-LOO. Performances are reported on dev sets for both datasets.

Granularity	Aspect R1	Aspect R2	Aspect RL	General R1	General R2	General RL
SPACE						
Sentence	33.11	10.98	27.59	40.26	12.04	23.42
Review	25.01	6.42	18.04	39.86	11.21	23.07
OPOSUM+						
Review	29.18	6.38	20.41	36.16	11.89	20.58
Sentence	22.34	5.06	17.33	20.06	6.76	13.86

Table 10. Pseudo summary granularity study for SW-LOO and NLI-LOO. Performances are reported on dev sets. Note that in our main experiments, we use sentence level pseudo summary for SPACE and review level for OPOSUM+

cally, we use `all-mpnet-base-v2`[4], which is a sentence embedding model finetuned on a 1B sentence pairs dataset with a self-supervised contrastive learning objective. Results in Table 9 show that there is no significant difference except general summarization for SPACE where using sentence embeddings is much worse than using aspect entailment probability.

K Pseudo Summary Granularity

We use different pseudo-summary granularity for two datasets: sentence level for SPACE and review level for OPOSUM+. Sentence level directly uses sampled filtered aspect-related review portion (SW-LOO) or sampled aspect-related review sentence (NLI-LOO) as pseudo-summary, and review level uses the original review that contains the sampled pseudo-summary as pseudo-summary. Results in Table 10 show the importance of design choices for synthetic datasets construction.

L T5 Model Sizes

We use different T5 sizes including T5-SMALL, T5-BASE, and T5-LARGE as summarization models. Results in Table 11 show that larger summarization models do not necessarily guarantee better

[4] https://www.sbert.net/docs/pretrained_models.html

	Model	SPACE R1	SPACE R2	SPACE RL	OPOSUM+ R1	OPOSUM+ R2	OPOSUM+ RL
Aspect Summary	SW-LOO						
	T5-SMALL	33.11	10.98	27.59	29.18	6.38	20.41
	T5-BASE	33.43	11.08	27.73	30.03	6.60	20.53
	T5-LARGE	33.70	10.77	27.60	28.98	6.15	20.32
	NLI-LOO						
	T5-SMALL	30.20	9.84	25.92	27.48	5.64	19.21
	T5-BASE	30.24	10.04	25.95	27.58	5.28	19.29
	T5-LARGE	30.61	9.50	25.68	27.14	5.47	19.55
General Summary	SW-LOO						
	T5-SMALL	40.26	12.04	23.42	36.16	11.89	20.58
	T5-BASE	41.31	12.47	23.12	35.53	11.65	20.33
	T5-LARGE	39.90	10.94	22.64	32.96	10.24	19.41
	NLI-LOO						
	T5-SMALL	41.17	12.34	25.13	31.10	10.09	19.32
	T5-BASE	37.49	11.44	22.91	26.51	6.74	17.08
	T5-LARGE	37.57	10.14	21.77	30.41	6.77	17.37

Table 11. Using different T5 sizes as summarization model. Performances are reported on dev sets for both datasets.

downstream ROUGE scores and sometimes even hurt downstream performances. Our hypothesis is that larger models overfit synthetic datasets and thus perform slightly worse on downstream evaluation sets.

SW-LOO Summaries	
Building	The pool area was very nice and the room was clean and comfortable.
Cleanliness	Our room was very clean and comfortable.
Food	The breakfast was great and the staff was very helpful and helpful.
Location	The hotel is located right next to the main road and is a short walk from the beach.
Rooms	The room was very clean and comfortable.
Service	The staff was very friendly and helpful.
General	The pool was very nice and clean. We were able to walk to the beach and Duval st. from the hotel, so we had a nice view of the harbor and the sea! The breakfast was great and we stayed in October and were very pleased with the location - right next to all the restaurants ... The room was small but very small and very comfortable with clean and comfortable beds.

NLI-LOO Summaries	
Building	The hotel is a beautiful old hotel.
Cleanliness	The room was clean and the staff was very helpful.
Food	The breakfast was great and the view from the rooftop was amazing.
Location	The location is great - just a short walk to the Spanish Steps and the metro station.
Rooms	The rooms are small by European standards, but very clean and comfortable.
Service	the service was excellent and the staff was very friendly and helpful.
General	The hotel is very clean and the staff is friendly and helpful. The room was very small and clean, but the bathroom was a bit small compared to the other rooms in the UK. It is OK to stay here again. I would stay there again if you want to go back to Europe! The location is great - the city is just ten minutes walk from the metro station andn't be disappointed with the price of the rooms.

Table 12. *General* and *aspect-level* summaries for a hotel in SPACE dataset generated by SW-LOO and NLI-LOO

SW-LOO Summaries	
Sound Quality	I love this headset. It's a great product, but it doesn't have any issues with the sound! It is OK if you are looking for something that can be used for your Samsung TV?
Comfort	I love this headset. It's a great headset for the price, but it doesn't fit my ear perfectly!
Ease of Use	I bought this for my Motorola. It is very easy to set up, and the buttons are very comfortable!
General	I haven't found any way of getting that to be consistently good. The earpieces are not as sturdy or high quality in material as a Motorola, but the buttons are quite accessible and the sound varies based on how it's fitting into my ears! The set is very comfortable and has great range (roughly 100 feet) and connects easily to my iPhone with me - but it is not too big for me to wear if it doesn't fit my TV.

NLI-LOO Summaries	
Sound Quality	I love these headphones. They are very comfortable, sound quality is good and they're very good quality for the price!
Comfort	I love these headphones. They are very comfortable, and the sound quality is great! They're a little tight on my ears but if you aren't sure how long they will last you...
Ease of Use	I bought these for my Motoactv. They are very comfortable to wear, and they don't touch my neck at all!
General	I bought these headphones in a package for the Motoactv. They are very comfortable, the neck band doesn't touch my neck at all allowing for free movement! The sound is very good and fits comfortably in my ears... but it takes some time to find the right angel and fit it right in.

Table 13. *General* and *aspect-level* summaries for a product in "Bluetooth Headset" domain of OPOSUM+ dataset generated by SW-LOO and NLI-LOO.