Unsupervised Opinion Summarization Using Approximate Geodesics

Anonymous ACL submission

Abstract

Opinion summarization is the task of creating summaries capturing popular opinions from user reviews. In this paper, we introduce Geodesic Summarizer (GeoSumm), a novel system to perform unsupervised extractive opinion summarization. GeoSumm involves an encoder-decoder based representation learning model, that generates representations of text as a distribution over latent semantic units. GeoSumm generates these representations by performing dictionary learning over pre-trained text representations at multiple layers of the decoder. We then use these representations to quantify the importance of review sentences using a novel approximate geodesic distance based scoring mechanism. We use the importance scores to identify popular opinions in order to compose general and aspect-specific summaries. Our proposed model, GeoSumm, achieves state-of-the-art performance on three opinion summarization datasets. We perform additional experiments to analyze the functioning of our model and showcase the generalization ability of GeoSumm across different domains.

1 Introduction

As more and more human interaction takes place online, consumers find themselves wading through an ever-increasing number of documents (e.g. customer reviews) when trying to make informed purchasing decisions. As this body of information grows, so too does the need for automatic systems that can summarize it in an unsupervised manner. Opinion summarization is the task of automatically generating concise summaries from online user reviews (Hu and Liu, 2004; Pang, 2008; Medhat et al., 2014). For instance, opinion summaries allow a consumer to understand product reviews without reading all of them. Opinion summaries are also useful for sellers to receive feedback, and compare different products. The recent success of deep learning techniques have led to a significant improvement in summarization (Rush et al., 2015; Nallapati et al., 2016; Cheng and Lapata, 2016; See et al., 2017; Narayan et al., 2018; Liu et al., 2018) in supervised settings. However, it is difficult to leverage these techniques for opinion summarization due to the scarcity of annotated data. It is expensive to collect good-quality opinion summaries as human annotators need to read hundreds of reviews to write a single summary (Moussa et al., 2018). Therefore, most works on opinion summarization tackle the problem in an unsupervised setting.

Recent works (Bražinskas et al., 2021; Amplayo et al., 2021a) focus on abstractive summarization, where fluent summaries are generated using novel phrases. However, these approaches suffer from issues like text hallucination (Rohrbach et al., 2018), which affects the faithfulness of generated summaries (Maynez et al., 2020). Extractive summaries are less prone to these problems, presenting the user with a representative subset of the original reviews.

We focus on the task of unsupervised extractive opinion summarization, where the system selects sentences representative of the user opinions. Inspired by previous works (Chowdhury et al., 2022; Angelidis et al., 2021a), we propose a novel encoder-decoder architecture along with objectives for (1) learning sentence representations that capture underlying semantics, and (2) a sentence selection algorithm to compose a summary.

One of the challenges in extractive summarization is quantifying the importance of opinions. An opinion is considered to be important if it is semantically similar to opinions from other users. Using off-the-shelf pre-trained representations to obtain semantic similarity scores has known issues (Timkey and van Schijndel, 2021). These similarity scores can behave counterintuitively due to the high anisotropy of the representation space (a few dimensions dominating the cosine similarity scores). Therefore, we use topical representations (Blei et al., 2003), which capture the underlying semantics of text as a distribution over latent semantic units, where the semantic units encode concepts or topics. These semantic units can be captured using a learnable dictionary (Engan et al., 1999; Mairal et al., 2009; Aharon et al., 2006; Lee et al., 2006). Topical representations enable us to effectively measure semantic similarity between text representations, as they are distributions over the same support. Text representations from reviews lie on a high-dimensional manifold. It is important to consider the underlying manifold while computing the importance score of a review. Therefore, we use the approximate geodesic distance between topical text representations to quantify the importance scores of reviews.

In this paper, we present **Geo**desic **Summ**arization (GeoSumm) that learns topical text representations in an unsupervised manner from distributed representations (Hinton, 1984). We also present a novel sentence selection scheme that compares topical sentence representations in high-dimensions using approximate geodesics. Empirical evaluations show that GeoSumm achieves state-of-the-art performance on three opinion summarization datasets – OPOSUM+ (Amplayo et al., 2021a), AMAZON (He and McAuley, 2016) and SPACE (Angelidis et al., 2021b). To summarize, our primary contributions are:

- We present an extractive opinion summarization system, GeoSumm. It consists of an unsupervised representation learning system and a sentence selection algorithm (Section 3).
- We present a novel representation learning model that learns topical text representations from distributed representations using dictionary learning (Section 3.1).
- We present a novel sentence selection algorithm that computes importance of text using approximate geodesic distance (Section 3.2).
- GeoSumm achieves state-of-the-art results on 3 opinion summarization datasets (Section 4.3).

2 Task Setup

In extractive opinion summarization, the objective is to select representative sentences from a reviews set. Specifically, each dataset consists of a set of entities $\mathbf{E}$ and their corresponding review set $\mathcal{R}$. For each entity $e \in \mathbf{E}$ (e.g., a particular hotel such as the Holiday Inn in Redwood City, CA.), a review set $\mathcal{R}_e = \{r_1, r_2, \ldots\}$ is provided, where each review is an ordered set of sentences $r_i = \{s_1^{(i)}, s_2^{(i)}, \ldots\}$. For simplicity of notation, we will represent the set of review sentences corresponding to an entity e as $\mathcal{S}_e = \bigcup_{r_i \in \mathcal{R}_e} r_i$. For each entity, reviews encompass a set of aspects $\mathcal{A}_e = \{a_1, a_2, \ldots\}$ (e.g., service, food of a hotel). In this work, we consider two forms of extractive summarization: (a) *general summarization*, where the system selects a subset of sentences $\mathcal{O}_e \subset \mathcal{S}_e$, that best represents popular opinions in the review set $\mathcal{R}_e$; (b) *aspect summarization*, where the system selects a representative sentence subset $\mathcal{O}_e^{(a)} \subset \mathcal{S}_e$, about a specific aspect a for entity e.

3 Geodesic Summarizer (GeoSumm)

In this section, we present our proposed approach Geodesic Summarizer (GeoSumm). GeoSumm has two parts: (a) an unsupervised model to learn topical representations of review sentences, and (b) a sentence selection algorithm, that uses approximate geodesic distance between topical representations, to compose the extractive summary.

3.1 Unsupervised Representation Learning

The goal of the representation learning model is to learn topical representations of review sentences. Topical representations model text as a distribution over underlying concepts or topics. This is useful for unsupervised extractive summarization because we want to capture the aggregate semantic distribution, and quantify the importance of individual review sentences with respect to the aggregate distribution. Topical representations allow us to achieve both. Being a distribution over latent units, topical representations can be combined to form an aggregate (mean) representation, enabling compositionality. Also, it is convenient to measure similarity between representations using conventional metrics (like cosine similarity).

We propose to model topical representations by decomposing pre-trained representations using dictionary learning (Tillmann, 2015; Lotfi and Vidyasagar, 2018). In this setup, the various components of the dictionary captures latent semantic units, and we consider the representation over dictionary elements as the topical representation. Unlike conventional dictionary learning algorithms, we use a sentence reconstruction objective for learning the dictionary. We use an encoder-decoder architecture to achieve this. We retrieve word embeddings from a pre-trained encoder. We modify

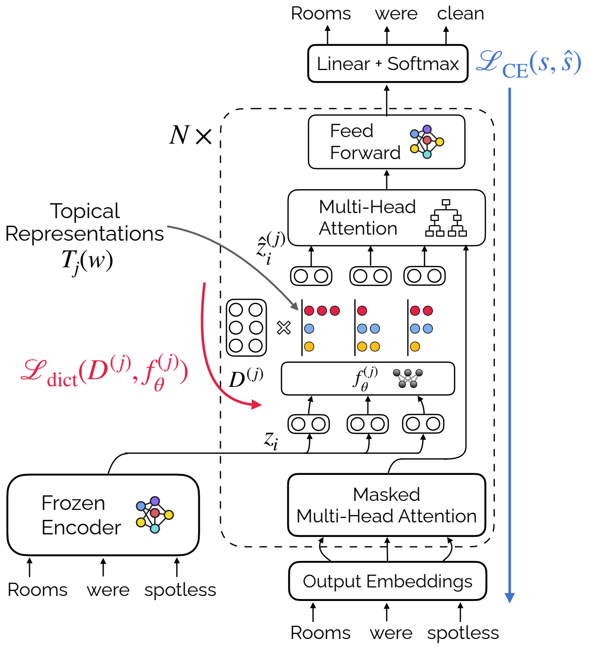

Figure 1: Architecture of Geodesic Summarizer. Sparse representations of words are formed via the kernel function $f_\theta^{(j)}$. The representations are trained to reconstruct the output embeddings of the encoder layer. Alongside the dictionary learning objective, we use an unsupervised sentence-reconstruction cross entropy loss. N indicates the number of decoder layers.

the architecture of a standard Transformer decoder, and add a dictionary learning component at each decoder layer. The pre-trained word embeddings obtained from the encoder are decomposed using these dictionary learning components to obtain topical representations. Then, we combine the topical word representations at different decoder layers to form a sentence representation. The schematic diagram of the model is shown in Figure 1. Next, we will discuss each of the components in detail.

Encoder. We obtain contextual word embeddings from a pre-trained BART (Lewis et al., 2020) encoder. We keep the weights of the encoder frozen during training.[1] Given an input sentence $s = \{w_1, \ldots, w_L\}$, we retrieve contextual word embeddings z_i's from the BART encoder

$$\mathrm{z}_i = \mathrm{sg}(\mathrm{enc}(\mathrm{w}_i)) \in \mathbb{R}^d \tag{1}$$

where $\mathrm{sg}(\cdot)$ denotes the stop gradient operator.

Dictionary Learning. We describe the dictionary learning component within each decoder layer. We use dictionary learning to decompose pre-trained word representations from the encoder to obtain a sparse representation for each word. We want word representations to be sparse because each word can capture only a small number of semantics. We forward word representations from the encoder to the decoder layers. For the j-th decoder layer, we use a dictionary, $\mathbf{D}^{(j)} \in \mathbb{R}^{m \times d}$, and kernel function, $k_j(\cdot, \cdot)$, where $j \in \{1, \ldots, N\}$ (N is the number of decoder layers). The dictionary captures the underlying semantics in the text by enabling us to model text representations as a combination of dictionary elements. Specifically, we learn a topical word representation $T_j(w_i)$ over the dictionary $\mathbf{D}^{(j)}$ as:

$$\begin{aligned} \hat{z}_i^{(j)} &= \mathbf{D}^{(j)^T} T_j(w_i) \\ T_j(w_i) &= k_j(\mathrm{z}_i, \mathbf{D}^{(j)}) \in \mathbb{R}^m \end{aligned} \tag{2}$$

where $\hat{z}_i^{(j)}$ is the reconstructed word embedding, and $k_j(\cdot, \cdot) \in \mathbb{R}^m$ is the kernel function that measures the similarity between z_i and individual dictionary elements. In practice, since the dictionary is common for all word embeddings z_i's, the kernel function can be implemented as:

$$k_j(\mathrm{z}_i, \mathbf{D}^{(j)}) = f_\theta^{(j)}(\mathrm{z}_i) \in \mathbb{R}^m \tag{3}$$

where $f_\theta^{(j)}$ is a feed-forward neural network with ReLU non-linearity. ReLU non-linearity ensures that the kernel coefficients are positive and also encourages sparsity.

Following conventional dictionary learning algorithms (Beck and Teboulle, 2009), the dictionary $\mathbf{D}^{(j)}$ and kernel layer $f_\theta^{(j)}$ are updated iteratively. This can be achieved by using the loss function:

$$\begin{aligned} \mathcal{L}_{\mathrm{dict}}(\mathbf{D}^{(j)}, f_\theta^{(j)}) = &||\mathrm{z}_i - \mathrm{sg}(\mathbf{D}^{(j)^T}) f_\theta^{(j)}(\mathrm{z}_i)||_2 + \\ &||\mathrm{z}_i - \mathbf{D}^{(j)^T} \mathrm{sg}(f_\theta^{(j)}(\mathrm{z}_i))||_2 \end{aligned}$$

where the gradient update of the dictionary $\mathbf{D}^{(j)}$ and kernel layer $f_\theta^{(j)}$ are performed independently.

Decoder. We build on the decoder architecture introduced by Vaswani et al. (2017). A decoder layer consists of 3 sub-layers (a) masked multi-head attention layer that takes as input decoder token embeddings, (b) multi-head attention that performs cross-attention between decoder tokens and encoder stack output, and (c) feed-forward network. We modify the cross attention multi-head sub-layer to attend over the reconstructed word embeddings $\hat{z}_i^{(j)}$ (Equation 2), instead of the encoder stack output (shown in Figure 1). Finally, the decoder autoregressively generates the reconstructed sentence $\hat{s} = \{\hat{w}_1, \ldots, \hat{w}_L\}$.

[1] In Section 5, we discuss why frozen representations are important for our model.

Training. The system is trained using the sentence reconstruction objective. The overall objective function is shown below:

$$\mathcal{L}_{\text{CE}}(s, \hat{s}) + \sum_{j=1}^{N} \mathcal{L}_{\text{dict}}(\mathbf{D}^{(j)}, f_\theta^{(j)}) \qquad (4)$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, and $f_\theta^{(j)}$ is the implementation of the kernel function $k_j(\cdot,\cdot)$ corresponding to the j-th decoder layer. The above loss function is used to update the decoder, dictionary elements and kernel parameters while keeping the encoder weights frozen.

Sentence Representations. We combine topical word representations from different decoder layers to form a sentence representation. First, we obtain a word representation, $T_j(w) \in \mathbb{R}^m$ from each decoder layer. We compose the final word representation $\mathbf{x}_w$ by concatenating representations from all decoder layers.

$$\mathbf{x}_w = [T_1(w), \ldots, T_N(w)] \in \mathbb{R}^{mN} \qquad (5)$$

where m is the dictionary dimension and N is the number of decoder layers. We use max-pooling over the dimensions of word representations to form a sentence representation $\mathbf{x}_s$ as shown below.

$$\begin{aligned} \mathbf{x}_n^s &= \max_{i\in\{1,\ldots,L\}} \mathbf{x}_w\big|_n \\ \bar{\mathbf{x}}_s &= \{\mathbf{x}_n^s\}_{n=1}^{mN}, \mathbf{x}_s = \bar{\mathbf{x}}_s/||\bar{\mathbf{x}}_s||_1 \in \mathbb{R}^{mN} \end{aligned} \qquad (6)$$

where $\mathbf{x}_w\big|_n$ is the n-th entry of the vector $\mathbf{x}_w$. The sentence representation $\mathbf{x}_s$ is normalized to a unit vector. Next, we discuss how we leverage these topical sentence representations to compute importance scores using approximate geodesics. We use the importance scores to compose the final extractive summary for a given entity.

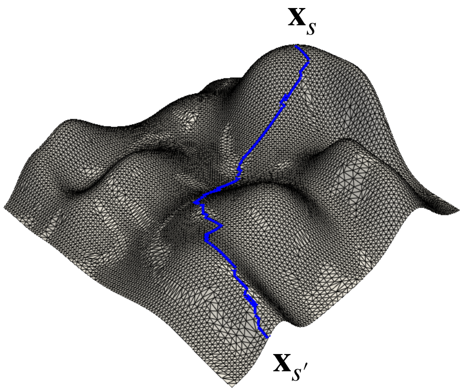

Figure 2: Illustration of the geodesic shortest path (shown in **blue**) between two sentence representations $\mathbf{x}_s$ and $\mathbf{x}_{s'}$ on a three-dimensional manifold.

3.2 General Summarization

We use representations retrieved from GeoSumm to select sentences representative of popular opinions in the review set. For an entity e, the set of sentence representations is denoted as $\mathcal{X}_e = \{\mathbf{x}_s | s \in \mathcal{S}_e\}$. For a summary budget q, we select a subset of sentences $\mathcal{O}_e \subset \mathcal{S}_e$ according to their importance scores, such that $|\mathcal{O}_e| = q$. First, we compute a mean representation as shown: $\mu_e = \mathbb{E}_{s\sim\mathcal{S}_e}[\mathbf{x}_s]$. Secondly, we define the importance of a sentence s, as the distance from the mean representation $d(\mathbf{x}_s, \mu_e)$. However, we do not directly evaluate $d(\cdot,\cdot)$ using a similarity metric. Representations in $\mathcal{X}_e$ lie in a high-dimensional manifold, and we aim to measure the geodesic distance (Jost and Jost, 2008) between two points along that manifold. An illustration of the geodesic distance between two points is shown in Figure 2. Computing the exact geodesic distance is difficult without explicit knowledge of the manifold structure (Surazhsky et al., 2005). We approximate the manifold structure using a k-NN graph. Each sentence representation forms a node in this graph. A directed edge exists between two nodes if the target node is among the k-nearest neighbours of the source node. The edge weight between two nodes (s, s') is defined using their cosine similarity distance, $d(s, s') = 1 - \mathbf{x}_s\mathbf{x}_{s'}^T$. The geodesic distance between two sentence representations is computed using the shortest path distance along the weighted graph. Therefore, the importance score $I(s)$ for a sentence s, is defined as:

$$I(s) = 1/\text{ShortestPath}(\mathbf{x}_s, \mu_e) \qquad (7)$$

where the shortest path distance is computed using Dijkstra's algorithm (Dijkstra et al., 1959). We select the top-q sentences according to their importance scores $I(s)$ to form the final general extractive summary. The overall sentence selection routine is shown in Algorithm 1.

3.3 Aspect Summarization

In aspect summarization, the goal is to select representative sentences to form a summary specific to an aspect (e.g., durability) of an entity (e.g., bag). To perform aspect summarization, we compute the mean representation of aspect-specific sentences as shown: $\mu_e^{(a)} = \mathbb{E}_{s\sim\mathcal{S}_e^{(a)}}[\mathbf{x}_s]$, where $\mathcal{S}_e^{(a)}$ is the

Algorithm 1 General Summarization Routine

```
Input: A set of sentence representations $\mathcal{X}_e = \{\mathbf{x}_s | s \in \mathcal{S}_e\}$ are review sentences for entity $e$.
$\mu_e \leftarrow \mathbb{E}_{s \sim \mathcal{S}_e}[\mathbf{x}_s]$
$\mathbf{A} \leftarrow \text{knn}(\mathcal{X}_e \cup \mu_e) \in \mathbb{R}^{l \times l}$    ▷ adjacency matrix of $k$-NN graph, $l = |S_e| + 1$.
$d \leftarrow \text{Dijkstra}(\mathbf{A}, \mu_e)$    ▷ shortest distances of all nodes from $\mu_e$
$I = \{1/d(s) | s \in \mathcal{S}_e\}$    ▷ importance scores
$t_q \leftarrow \min \text{top-}q(I)$    ▷ top-$q$ threshold
$\mathcal{O}_e \leftarrow \{s \mid I(s) \geq t_q, s \in \mathcal{S}_e\}$
return $\mathcal{O}_e$
```

set of sentences mentioning aspect a. We identify $\mathcal{S}_e^{(a)}$ by detecting the presence of aspect-specific keywords available with the dataset. To ensure the selected sentences are aspect-specific, we introduce a measure of *informativeness* (Chowdhury et al., 2022; Peyrard, 2019). Informativeness penalizes a sentence for being close to the overall mean μ_e. Therefore, we model the aspect-specific importance score $I_a(s)$ as:

$$I_a(s) = 1/\text{ShortestPath}(\mathbf{x}_s, \mu_e^{(a)}) - \gamma I(s) \quad (8)$$

where γ is a hyperparameter, $I(s)$ is the overall importance score (obtained from Eqn. 7). Aspect summary $\mathcal{O}_e^{(a)}$ is composed using the top-q sentences according to the aspect-specific scores, $I_a(s)$.

4 Experiments

We evaluate the performance of GeoSumm on extractive summarization. Given a set of opinion reviews the system needs to select a subset of the sentences as the summary. This summary is then compared with human-written summaries. In this section, we discuss the experimental setup in detail.

4.1 Datasets & Metrics

We evaluate GeoSumm on three publicly available opinion summarization datasets:

(a) OPOSUM+ (Amplayo et al., 2021b) is an extended version of the original OPOSUM dataset (Angelidis and Lapata, 2018a). This dataset contains Amazon reviews from six product categories (like laptops, bags, etc.), with 3 human-written summaries in the test set. The extended version contains additional product reviews and aspect-specific human-annotations.

(b) AMAZON (He and McAuley, 2016; Bražinskas et al., 2020a) has product reviews of 4 different categories (like electronics, clothing etc.) from Amazon, with 3 human summaries per entity.

Dataset	Reviews	Train / Test Ent.	Rev./Ent.
OPOSUM+	4.13M	95K /60	10
AMAZON	4.75M	183K / 60	8
SPACE	1.14M	11.4K / 50	100

Table 1: Dataset statistics for OPOSUM+, AMAZON and SPACE datasets. (Train/Test Ent.: Number of entities in the *training* and *test* set; Rev./Ent.: Number of reviews per entity in the *test* set.)

(c) SPACE (Angelidis et al., 2021a) contains reviews for hotels from Tripadvisor. SPACE provides 3 human-written abstractive summaries per entity. It also has 6 aspect-specific summaries per entity.

Statistics of the datasets are provided in Table 1. We observe that SPACE dataset has significantly more reviews per entity compared to other datasets.

4.2 Baselines

We compare GeoSumm with several summarization systems (including the current state-of-the-art) that can be classified into three broad categories:

- *Single Review* systems select a single review as the summary. We compare with the following systems: (a) *Random* samples a review randomly from the review set; (b) *Centroid* selects a review closest to the centroid of the review set. The centroid is computed using BERT (Devlin et al., 2019) embeddings; (c) *Oracle* selects the best review based on ROUGE overlap with the human-written summary.
- *Abstractive* systems generate summaries using novel phrasing. We compare GeoSumm with the following systems: MeanSum (Chu and Liu, 2019), Copycat (Bražinskas et al., 2020b), and AceSum (Amplayo et al., 2021b).
- *Extractive* systems select text phrases from the review set to form the summary. We compare with the following systems: LexRank (Erkan and Radev, 2004) using BERT embeddings, QT (Angelidis et al., 2021a), AceSum$_{\text{EXT}}$ (Amplayo et al., 2021b), and SemAE[2] (Chowdhury et al., 2022).

4.3 Results

We discuss the performance of GeoSumm on general and aspect-specific summarization. We evaluate the quality of the extracted summaries using the automatic metric – ROUGE F-scores (Lin, 2004).

General Summarization. We present the results of GeoSumm and baseline approaches on general

[2] For a fair comparison, we consider the version of SemAE that does not use additional aspect-related information.

	Method	OpoSum+			Amazon			Space		
		R1	R2	RL	R1	R2	RL	R1	R2	RL
Single Rev.	Random	29.88	5.64	17.19	27.66	4.72	16.95	26.24	3.58	14.72
	Centroid$_{\text{BERT}}$	33.44	11.00	20.54	29.94	5.19	17.70	31.29	4.91	16.43
	Oracle	32.89	23.20	28.73	31.69	6.47	19.25	33.21	8.33	18.02
Abstract	MeanSum (Chu and Liu, 2019)	34.95	7.49	19.92	29.20	4.70	18.15	34.95	7.49	19.92
	Copycat (Bražinskas et al., 2020b)	36.66	8.87	20.90	31.97	5.81	20.16	36.66	8.87	20.90
	AceSum (Amplayo et al., 2021c)	40.37	11.51	23.23	-	-	-	40.37	11.51	23.23
Extract	LexRank$_{\text{BERT}}$ (Erkan and Radev, 2004)	35.42	10.22	20.92	31.47	5.07	16.81	31.41	5.05	18.12
	QT (Angelidis et al., 2021a)	37.72	14.65	21.69	31.27	5.03	16.42	38.66	10.22	21.90
	AceSum$_{\text{EXT}}$ (Amplayo et al., 2021b)	38.48	15.17	22.82	-	-	-	35.50	7.82	20.09
	SemAE (Chowdhury et al., 2022)	39.16	16.85	23.61	32.03	5.38	16.47	42.48	**13.48**	26.40
	Geodesic Summarizer (GeoSumm)	**41.29**	**19.94**	**33.53**	**32.93**	**6.91**	**25.45**	**43.29**	12.80	**29.87**

Table 2: Evaluation results of GeoSumm and baseline approaches on general summarization. We observe that GeoSumm achieves state-of-the-art performance on all datasets. GeoSumm significantly improves ROUGE-L scores, with an average improvement of 6.9 points over prior best. ROUGE-L being the most difficult metric of overlap, showcases the efficacy of GeoSumm in selecting sentences that correlate with human summaries. We report the ROUGE-F scores denoted as – R1: ROUGE-1, R2: ROUGE-2, RL: ROUGE-L.

	Method	OpoSum+			Space		
		R1	R2	RL	R1	R2	RL
Abstract	MeanSum	24.63	3.47	17.53	23.24	3.72	17.02
	CopyCat	26.17	4.30	18.20	24.95	4.82	17.53
	AceSum	29.53	6.79	21.06	**32.41**	9.47	**25.46**
Extract	LexRank	22.51	3.35	17.27	27.72	7.54	20.82
	QT	23.99	4.36	16.61	28.95	8.34	21.77
	SemAE	25.30	5.08	17.62	31.24	**10.43**	24.14
	AceSum$_{\text{EXT}}$	26.16	5.75	18.55	30.91	8.77	23.61
	GeoSumm	**30.64**	**7.94**	**24.37**	30.29	9.02	23.79

Table 3: Evaluation results on aspect summarization. Best scores for each metric is highlighted in **bold**. GeoSumm achieves the state-of-the-art performance on OpoSum+, while achieving competitive performance with other extractive methods on Space.

summarization in Table 2. We observe that GeoSumm achieves strong improvement over baselines (including abstractive summarization approaches) across all datasets. It is important to note that compared to baselines GeoSumm achieves much better ROUGE-L F1 scores, which is the hardest metric among the three (with an average improvement of 6.9 points over prior best). This shows the efficacy of GeoSumm in selecting sentences that correlate with human summaries. For Space dataset, it is competitive with the state-of-the-art model SemAE, falling slightly short only in ROUGE-2 F1 score. However, GeoSumm performs significantly better than SemAE on human evaluations.

Aspect Summarization. We report the performance on different approaches on aspect summarization in Table 3 on OpoSum+ and Space. We observe that GeoSumm achieves the state-of-the-art performance for all metrics on OpoSum+ dataset. It also achieves strong results on Space, obtaining similar scores compared to other extractive approaches, falling slightly short of state-of-the-art.

General	Inform.	Coherence	Redund.
SemAE	-29.3	-25.3	-58.0
QT	4.0	-19.3	**40.7**
GeoSumm	**25.3**	**44.7***	17.3

Table 4: Human evaluation results of general summarization for Space dataset. (*): statistically significant difference with all baselines ($p < 0.05$, using paired bootstrap resampling Koehn (2004)).

Human Evaluation. We perform human evaluation to compare the summaries from GeoSumm with the state-of-the-art extractive summarization systems SemAE and QT. General summaries were judged based on the following criteria: *informativeness*, *coherence* and *redundancy*. We present human evaluators with summaries in a pairwise fashion, and ask them to select which one was better/worse/similar according to the criteria. The final scores for each system reported in Table 4 were computed using Best-Worst Scaling (Louviere et al., 2015). We observe that GeoSumm outperforms the prior state-of-the-art in informativeness and coherence. GeoSumm performs slightly worse than QT in redundancy. This is expected as GeoSumm greedily select sentences, while QT performs sampling to introduce diversity in the summary (compromising on informativeness).

For aspect summaries, we ask annotators to judge whether a summary discusses a specific aspect *exclusively*, *partially*, or *does not mention* it

Aspect	Exclusive	Partial	None
SemAE	22.1	43.8	34.1
QT	22.2	41.9	35.9
GeoSumm	**45.7***	**40.1**	**14.1***

Table 5: Human evaluation results of aspect summarization for OPOSUM+ dataset. GeoSumm generates more aspect-specific summaries compared to baselines.

at all. In Table 5, we report the human evaluation results for aspect summaries on OPOSUM+ dataset. We observe that GeoSumm generates summaries that are more specific to an aspect compared to baselines. We provide further details about human evaluation and additional results in Appendix A.3.

5 Analysis

Thawed Encoder. In this experiment, we compare the performance of GeoSumm when the encoder is allowed to be fine-tuned with the original setup, where the encoder weights are frozen. In Table 6, we observe that there is a significant drop in performance when the encoder is fine-tuned. We believe that this happens because the model overfits on shallow word-level semantics, and is unable to capture more abstract semantics. This showcases the utility of pre-trained representations, which helps GeoSumm perform well in an unsupervised setting.

Dataset	R1	R2	RL
OPOSUM+	31.46 [-9.83]	8.60 [-11.34]	23.94 [-9.59]
AMAZON	30.12 [-2.81]	4.85 [-2.06]	22.01 [-3.44]
SPACE	30.07 [-13.22]	4.31 [-8.49]	21.56 [-8.31]

Table 6: Evaluation results of GeoSumm when the encoder is fine-tuned during training.

Visualization. In this experiment, we visualize the UMAP (McInnes et al., 2018) projections of sentence representation retrieved from GeoSumm for an entity. We investigate whether different parts of the representation space capture distinct semantics. We partition the space using kmeans clustering ($k = 10$) on the representations, color-code them according to the assigned cluster label and visualize them in Figure 3. In Table 7, we observe that these clusters capture certain semantics. We report example sentences within different clusters. We observe that sentences belonging to the same cluster share a common theme. The underlying semantics of a cluster can vary from being coarse, like presence of a phrase 'Calistoga', to more nuanced concepts like cleanliness of rooms, time frame etc.

Next, we investigate the efficacy of the representation learning and sentence selection modules by replacing each of them with a competitive variant.

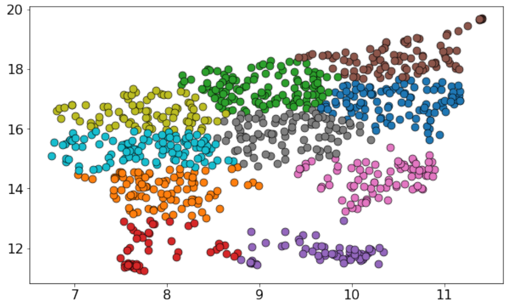

Figure 3: UMAP projections of sentence representation retrieved from GeoSumm for an entity. The representations are colored-coded according to the cluster labels.

Theme	Sentences
Cleanliness ●	• The room was clean, but no more than that • Also, the pools were filthy dirty.
Time ●	• I called on Monday, and was told there was no manager on Mondays! • save your time and money and stay anywhere else.
Location 'Calistoga' ●	• The Roman Spa and Calistoga is our favorite spot in the Wine Country. • Roman Spa Hot Springs Resort in Calistoga is a wonderful place...
Pleasant Rooms ●	• The rooms were in great shape, very clean, comfortable beds ... • The rooms are very comfortable and they have upgraded them ...

Table 7: Sentences within a cluster visualized in Figure 3. Sentences in a row belong to the same cluster. We highlight the dominant theme of a cluster in green. We annotate cluster identity using color-coded circles.

Euclidean-based Importance Score. We investigate the utility of geodesic-based importance scoring over Euclidean-based scoring. In this experiment, instead of $I(s)$ (defined in Equation 7) we compute the importance score of a sentence (s) as the Euclidean distance from the mean representation, μ_e ($I(s) = -||\mathbf{x}_s - \mu_e||_2^2$). We report the results of this setup in Table 8 (relative performance to GeoSumm is shown in brackets). We observe that using Euclidean distance achieves similar performance for AMAZON and OPOSUM+ (with a slight improvement for AMAZON). But there is a significant performance drop for SPACE. SPACE has a larger number of reviews per entity, providing a better approximation of the manifold using k-NN graph, and therefore a more accurate geodesic distance. We believe that this is the reason why Euclidean distance achieves comparable performance, when there are less reviews. In the real word, opinion summarization involves a large number of reviews, where GeoSumm will scale better.

Distributed vs. Topical Representations. In this

Dataset	R1	R2	RL
OPOSUM+	40.87 [-0.42]	19.66 [-0.28]	32.85 [-0.68]
AMAZON	33.42 [+0.49]	6.95 [+0.04]	25.72 [+0.27]
SPACE	40.95 [-2.34]	11.21 [-1.59]	28.57 [-1.30]

Table 8: Evaluation results of GeoSumm with a modified score $I(s) = -||\mathbf{x}_s - \mu_e||_2^2$. We observe a significant drop in performance on SPACE, while achieving similar performance on OPOSUM+ and AMAZON.

experiment, we investigate the relative efficacy of topical representations compared to distributed representations. We retrieve distributed sentence representations from RoBERTa (Liu et al., 2019) (`[CLS]` token feature) and SimCSE (Gao et al., 2021) model. Then, we use our sentence selection algorithm (Section 3.2) to compose the summary. In Table 9, we observe that topical representations outperform distributed representations by a significant margin in almost all setups (except AMAZON with RoBERTa embeddings). This shows the utility of topical representations over distributed representations for unsupervised summarization.

We perform additional experiments to investigate the domain transfer capabilities, sparsity of representation, and visualize the generated summaries from GeoSumm in Appendix A.2.

6 Related Work

Most work on opinion summarization focuses on generating summaries in an unsupervised setup due to scarcity of labeled data. These works are broadly classified into two categories based on the type of summaries being generated: *abstractive* (Ganesan et al., 2010; Carenini et al., 2006; Di Fabbrizio et al., 2014) or *extractive* (Erkan and Radev, 2004; Nenkova and Vanderwende, 2005; Kim et al., 2011). Abstractive systems, in an unsupervised setup (Chu and Liu, 2019; Bražinskas et al., 2020b; Iso et al., 2021; Wang and Wan, 2021; Amplayo et al., 2021a) train an encoder-decoder setup using a self-supervised objective, and generate the summary by leveraging the aggregate opinion representation. On the other hand, extractive opinion systems (Kim et al., 2011), select sentences using an importance score that quantifies its salience. Salience has been computed using frequency-based approaches (Nenkova and Vanderwende, 2005), distance from mean (Radev et al., 2004), or graph-based techniques (Erkan and Radev, 2004). Few approaches focus on aspect specificity and sentiment polarity for sentence selection (Angelidis and Lapata, 2018b; Zhao and Chaturvedi, 2020).

Dataset	Model	R1	R2	RL
OPOSUM+	RoBERTa	36.29 [-6.00]	13.18 [-7.76]	28.36 [-5.17]
	SimCSE	35.37 [-6.92]	12.99 [-6.95]	20.76 [-12.77]
AMAZON	RoBERTa	33.79 [+0.86]	6.95 [+0.04]	25.54 [+0.09]
	SimCSE	32.71 [-0.22]	6.53 [-0.38]	17.65 [-6.80]
SPACE	RoBERTa	38.70 [-4.59]	9.40 [-3.40]	27.60 [-2.27]
	SimCSE	35.36 [-7.93]	7.21 [-5.59]	19.72 [-10.15]

Table 9: Evaluation results of GeoSumm using RoBERTa and SimCSE's representations. We observe a significant drop in performance in most setups.

Our work is most similar to extractive summarization systems SemAE (Chowdhury et al., 2022), and QT (Angelidis et al., 2021a). Similar to these systems, Geodesic Summarizer has two components: a representation learning system, and a sentence selection routine. However, unlike these approaches, we leverage pre-trained models to learn topical representations over a latent dictionary, and propose a sentence selection mechanism using approximate geodesics to perform summarization.

Prior work in deep clustering consider a similar combination of unsupervised representation learning and sparse structures (Yang et al., 2016; Jiang et al., 2016; Law et al., 2017; Caron et al., 2020; Zhao et al., 2020; Chan et al., 2022). Similarly, dictionary learning-like approaches have been combined with deep networks (Liang et al., 2021; Zheng et al., 2021) for various tasks.

7 Conclusion

We present Geodesic Summarizer, a novel framework for extractive opinion summarization. GeoSumm uses a representation learning model to convert distributed representations from a pre-trained model into topical text representations. GeoSumm uses these representations to compute the importance of a sentence using approximate geodesics. We show that GeoSumm achieves state-of-the-art results on several opinion summarization datasets. However, there are a lot of open questions about the inductive biases of representation learning that are needed for unsupervised summarization. In this work, we show the efficacy of topical representations. However, are there better approaches to capture language semantics that help us quantify the importance of an opinion? Our analysis shows that representations from GeoSumm span the high-dimensional space in a manner that different parts of it capture distinct semantics. This opens up the possibility of leveraging the representation geometry to capture different forms of semantics. Future work can explore ways to leverage topical representations from GeoSumm for tasks where there is a scarcity of labeled data.

Ethical Considerations

We do not foresee any ethical issues from the technology introduced in this paper. However, we would like to mention certain limitations of extractive summarization systems in general. As extractive systems select review sentences from the input, it can produce undesirable output when the input reviews have foul or offensive language. Therefore, it is important to remove foul language from the input in order to ensure the end user is not affected. In general, we use public datasets, and do not annotate any data manually. All datasets used in this paper have customer reviews in English language. Human evaluations for summarization were performed on Amazon Mechanical Turks (AMT) platform. Human judges were based in the United States. Human judges on AMT were compensated at a wage rate of at least $15 USD per hour.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Michal Aharon, Michael Elad, and Alfred Bruckstein. 2006. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing.

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. Aspect-controllable opinion summarization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6578–6593.

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

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021a. Extractive opinion summarization in quantized transformer spaces. Transactions of the Association for Computational Linguistics, 9:277–293.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021b. Extractive opinion summarization in quantized transformer spaces. Transactions of the Association for Computational Linguistics, 9:277–293.

Stefanos Angelidis and Mirella Lapata. 2018a. 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.

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.

Amir Beck and Marc Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1):183–202.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

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

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

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

Giuseppe Carenini, Raymond Ng, and Adam Pauls. 2006. Multi-document summarization of evaluative text. In 11th Conference of the European Chapter of the Association for Computational Linguistics, pages 305–312.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems, 33:9912–9924.

Kwan Ho Ryan Chan, YD Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. 2022. Redunet: A white-box deep network from the principle of maximizing rate reduction. J Mach Learn Res, 23(114):1–103.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 484–494, Berlin, Germany. 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.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 1223–1232. PMLR.

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.

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.

Edsger W Dijkstra et al. 1959. A note on two problems in connexion with graphs. Numerische mathematik, 1(1):269–271.

Kjersti Engan, Sven Ole Aase, and John Håkon Husøy. 1999. Method of optimal directions for frame design. International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP).

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

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In proceedings of the 25th international conference on world wide web, pages 507–517.

Geoffrey E Hinton. 1984. Distributed representations.

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.

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.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. 2016. Variational deep embedding: An unsupervised and generative approach to clustering. arXiv preprint arXiv:1611.05148.

Jürgen Jost and Jeurgen Jost. 2008. Riemannian geometry and geometric analysis, volume 42005. Springer.

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

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

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Marc T Law, Raquel Urtasun, and Richard S Zemel. 2017. Deep spectral clustering learning. In International conference on machine learning, pages 1985–1994. PMLR.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. 2006. Efficient sparse coding algorithms. Advances in neural information processing systems, 19.

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.

Paul Pu Liang, Manzil Zaheer, Yuan Wang, and Amr Ahmed. 2021. Anchor & transform: Learning sparse embeddings for large vocabularies. International Conference on Learning Representations (ICLR).

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.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

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.

Mahsa Lotfi and Mathukumalli Vidyasagar. 2018. A fast noniterative algorithm for compressive sensing using binary measurement matrices. IEEE Transactions on Signal Processing, 66(15):4079–4089.

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

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2009. Online dictionary learning for sparse coding. In International Conference on Machine Learning (ICML).

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

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. Journal of Open Source Software, 3(29):861.

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

Mohammed Elsaid Moussa, Ensaf Hussein Mohamed, and Mohamed Hassan Haggag. 2018. A survey on opinion summarization techniques for social media. Future Computing and Informatics Journal, 3(1):82–109.

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

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

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005, 101.

Bo Pang. 2008. lee, l.(2008). opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2):1–135.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

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

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

Vitaly Surazhsky, Tatiana Surazhsky, Danil Kirsanov, Steven J Gortler, and Hugues Hoppe. 2005. Fast exact and approximate geodesics on meshes. ACM transactions on graphics (TOG), 24(3):553–560.

Andreas M. Tillmann. 2015. On the computational intractability of exact and approximate dictionary learning. IEEE Signal Processing Letters, 22(1):45–49.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4527–4546, Online and Punta Cana, Dominican Republic. 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 Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Ke Wang and Xiaojun Wan. 2021. TransSum: Translating aspect and sentiment embeddings for self-supervised opinion summarization. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 729–742, Online. Association for Computational Linguistics.

Jianwei Yang, Devi Parikh, and Dhruv Batra. 2016. Joint unsupervised learning of deep representations and image clusters. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5147–5156.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9644–9651.

Jinyu Zhao, Yi Hao, and Cyrus Rashtchian. 2020. Unsupervised embedding of hierarchical structure in euclidean space. arXiv preprint arXiv:2010.16055.

Hongyi Zheng, Hongwei Yong, and Lei Zhang. 2021. Deep convolutional dictionary learning for image denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

A Appendix

A.1 Implementation Details

Our experiments are implemented in TensorFlow (Abadi et al., 2015) framework. We use BART (Lewis et al., 2020) architecture as our encoder-decoder model. We initialize the encoder with pre-trained weights from BART, while the decoder is trained from scratch. In our experiments, we use dictionary dimension $m = 8192$, number of decoder layers $N = 6$, and hidden dimension $d = 768$. GeoSumm was trained for 15K steps on 16 TPUs in all setups. We optimize our model using Adam (Kingma and Ba, 2014) optimizer with a learning rate of 10^{-5}. We set aspect-summarization parameter $\gamma = 0.5$ for OpoSum+ and $\gamma = 0.7$ for Space (Equation 8). All hyperparameters were tuned using grid-search on the development set. We will make our code publicly available.

A.2 Analysis

Dictionary Size Ablation. In this experiment, we vary the number of elements in each dictionary (n) and observe the summarization performance on OpoSum+ dataset. We conduct these experiments on the OpoSum+ dataset. In Table 10, we observe GeoSumm achieves comparable performance with significantly smaller dictionary sizes.

m	R1	R2	RL
512	39.52	18.13	31.78
1024	40.03	19.14	32.69
2048	40.15	19.26	32.93
4096	**41.29**	**19.94**	**33.53**

Table 10: Evaluation results with varying number of dictionary elements on OpoSum+ dataset. We observe that there is only a small drop in performance of GeoSumm, when the dictionary sizes are reduced.

Sparsity. We investigate whether word representations from GeoSumm are sparse. We compute the number of non-zero elements in each word representation. We plot the histogram corresponding to the number of non-zero elements in representations. In Figure 4, we observe that the histogram is left-skewed which shows that most representations have a small number of non-zero elements. This shows that word representations are modeled as a combination of small number of latent semantics.

Domain Transfer capability. In this experiment, we investigate the domain transfer capability of GeoSumm. Specifically, we evaluate how GeoSumm trained on one dataset, performs on others. We also evaluate GeoSumm when it is trained on C4 dataset (Raffel et al., 2020). In Table 11, we report the results for this experiment. When evaluated on OpoSum+ or Amazon, we observe that GeoSumm is generalizing well, out-of-domain performance is comparable or better than in-domain performance (highlighted in gray). When evaluated on Space, we observe the out-of-domain performance to be much lower than in-domain performance. We also observe that the performance is the worst compared to others when tranferring from Space to other datasets. We hypothesize that this happens due to a domain shift, where both Amazon and OpoSum+ are product review datasets, while Space has reviews for hotel entities.

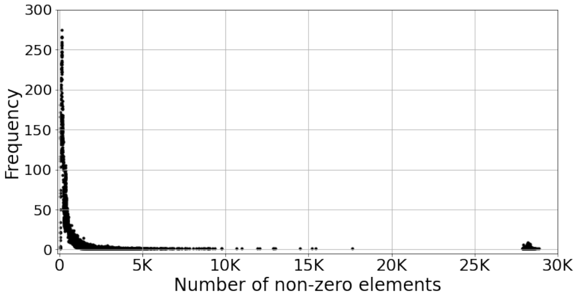

Figure 4: Plot depicting the sparsity of word representations retrieved from GeoSumm. We compute the number of non-zero elements in each word representation, and plot the corresponding histogram. We observe that the histogram is left-skewed showcasing sparsity among the representations.

Train→Predict	R1	R2	RL
Space→OpoSum+	39.06	17.48	31.09
Amazon→OpoSum+	39.14	18.44	31.77
C4→OpoSum+	**42.83**	**21.21**	**34.86**
OpoSum+→OpoSum+	41.29	19.94	33.53
Space→Amazon	32.10	6.52	24.68
OpoSum+→Amazon	32.34	6.62	24.78
C4→Amazon	**33.39**	6.88	**25.54**
Amazon→Amazon	32.93	**6.91**	25.45
OpoSum+→Space	33.53	7.48	24.43
Amazon→Space	36.30	9.26	25.72
C4→Space	32.55	6.46	24.07
Space→Space	**43.29**	**12.80**	**29.87**

Table 11: Evaluation results when the representation learning system is trained on a different dataset. In-domain performance is highlighted in gray. GeoSumm shows decent domain transfer performance for OpoSum+ and Amazon datasets. However,

Generated Summaries. In Table 12, we report the

Human	GeoSumm	SemAE	QT
All staff members were friendly, accommodating, and helpful. The hotel and room were very clean. The room had modern charm and was nicely remodeled. The beds are extremely comfortable. The rooms are quite with wonderful beach views. The food at Hash, the restaurant in lobby, was fabulous. The location is great, very close to the beach. It's a longish walk to Santa Monica. The price is very affordable.	The Hotel is classy and has a rooftop bar. The food and service at the restaurant was awesome. We ate breakfast at the hotel and it was great. Overall we had a nice stay at the hotel. I appreciate the location and the security in the hotel. The location is very central. It is very close to ocean, the stuff is friendly, rooms are clean. Our room was very clean and comfortable. It was great.	The staff is great. The Hotel Erwin is a great place to stay. The staff were friendly and helpful. The location is perfect. We ate breakfast at the hotel and it was great. The hotel itself is in a great location. The service was wonderful. It was great. The rooms are great. The rooftop bar HIGH was the icing on the cake. The food and service at the restaurant was awesome. The service was excellent.	Great hotel. We liked our room with an ocean view. The staff were friendly and helpful. There was no balcony. The location is perfect. Our room was very quiet. I would definitely stay here again. You're one block from the beach. So it must be good! Filthy hallways. Unvacuumed room. Pricy, but well worth it.

Table 12: Human-written and generated summaries from GeoSumm, SemAE, and QT. For fair comparison, we present the summary for the instance reported by in previous works. GeoSumm generates summaries where sentences with similar aspects appear together (highlighted in **green**), without abrupt context switch between aspects as seen in summaries of other approaches (highlighted in **red**).

General	Inform.	Coherence	Redund.
SemAE	18.9	-13.3	-16.1
AceSum	-53.9	**7.2**	**21.7**
GeoSumm	**35.0**	6.1	-5.6

Table 13: Human evaluation results of general summarization for OPOSUM+ dataset. We observe that GeoSumm generates the most informative summaries, while falling slightly behind the abstractive baseline (AceSum) in coherence and redundancy.

summaries generated by GeoSumm, and other comparable extractive summarization systems like SemAE and QT. We observe that GeoSumm is able to generate summaries where sentences with similar aspects stay together (highlighted in green) while covering multiple aspects of an entity. This shows that GeoSumm's representation learning system is able to capture underlying aspects, and we can effectively quantify them using geodesic distance. Baseline methods SemAE and QT, also cover different aspect but the summary abrupted switches between aspects (highlighted in red).

A.3 Human Evaluation

We perform human evaluation on the Amazon Mechanical Turk (AMT) platform. We designed the payment rate per Human Intelligence Task (HIT) in a manner to ensure that judges were compensated at a rate of at least $15 USD per hour. In all tasks, each HIT was evaluated by 3 human judges.

For general summarization, we performed pairwise evaluation of two summarization systems. Specifically, we given two system summaries the human judges were asked to judge each pair as better, worse or similar. We asked the judges to evaluate pair based on the following criteria – *informativeness*, *redundancy* and *coherence*, in independent tasks. For informativeness, we also provide the judges with a human-written summary. The judges annotate a summary as more informative only if the information is consistent with the human-written summaries. The reported scores (-100 to +100) were computed using Best-worst scaling (Louviere et al., 2015).

For aspect summarization, we provide human judges with a system generated aspect-summary and the corresponding aspect. Judges were asked to annotate whether the system summary discusses the mentioned aspect *exclusively*, *partially* or *does not mention* the aspect at all. In this setup, each system was evaluated individually by 3 human judges.

We present the human evaluation results of general summarization in Table 13. We compare GeoSumm with state-of-the-art extractive baseline SemAE and abstractive baseline AceSum. We observe that GeoSumm generates the most informative summaries compared to the baselines. It slightly falls short in coherence and redundancy when compared to the abstractive baseline – AceSum, which is expected because abstractive systems can generate summaries using novel phrases to ensure coherence. In extractive summarization, we focus on selecting relevant sentences without

Aspect	Exclusive	Partial	None
AceSum	50.9	42.6	6.5
GeoSumm	57.7	33.8	8.5
SemAE	**70.7***	**25.9***	**3.4***

Table 14: Human evaluation results of aspect summarization for SPACE dataset. (*): statistically significant difference with all baselines ($p < 0.05$, using paired bootstrap resampling Koehn (2004)).

considering the coherence of the generated summary. We still find that GeoSumm is competitive with abstractive baselines, which shows the efficacy of our approach.

Next, we present the human evaluation results of aspect summarization on SPACE dataset. We observe that SemAE generates the most aspect-specific summaries for this dataset. To investigate further why GeoSumm falls behind SemAE on SPACE, we analyze the performance of the systems for each aspect category. For each aspect category, we report the percentage of summaries annotated as exclusively aspect-specific for SemAE and GeoSumm (shown in Figure 5).

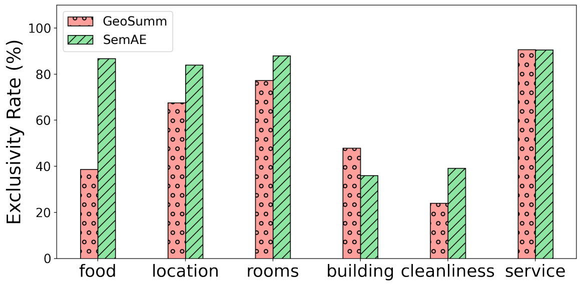

Figure 5: Plot showcasing the percentage of summaries annotated as exclusively aspect-specific for each aspect in the SPACE dataset.

We observe that SemAE achieves better or similar performance than GeoSumm for all aspect categories, except "building". In SPACE dataset, most aspects can be identified by the presence of seeds words, e.g., service – has words "staff", cleanliness – "clean", "spotless", rooms – "room", etc. However, the "building" aspect covers a variety of things like decor, pool, lounge, etc. We hypothesize that SemAE overfits on the word-level semantics, and just selects sentences that have lexical overlap with the seed sentences. This works in SPACE dataset because most aspect-specific sentences happen to contain a small number of words. On the other hand, GeoSumm is captures semantic information using the pre-trained encoder. This helps GeoSumm achieve good performance on aspects like "building", where selecting sentences based on the presence of a word is not helpful.

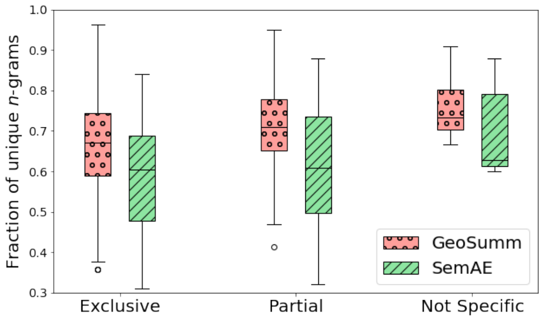

Figure 6: Plot showcasing the fraction of unique n-grams in a summary for each annotation label. We observe that GeoSumm generates more diverse summaries compared to SemAE.

We also observe that aspect summaries from SemAE are quite redundant. This helps SemAE generate summaries that are aspect-specific and achieve good ROUGE scores. However, they are not informative to the user. To quantify redundancy, we compute the fraction of unique n-grams ($n = 1$) in a summary. In Figure 6, we report the variation of the fraction of unique n-grams for each annotated category – *exclusive*, *partial*, and *not specific*. We observe a general trend that aspect summaries that are partially or not aspect-specific tend to be more diverse. We also observe GeoSumm generates more diverse summaries than SemAE across all annotation labels. Moreover, summaries from GeoSumm are more diverse than SemAE's summaries even in the exclusively aspect-specific category. This shows the efficacy of GeoSumm in generating more diverse and informative aspect summaries compared to baselines.