Hierarchical Indexing for Retrieval-Augmented Opinion Summarization

Tom Hosking **Hao Tang** **Mirella Lapata**
Institute for Language, Cognition and Computation,
School of Informatics, University of Edinburgh,
10 Crichton Street, Edinburgh EH8 9AB, United Kingdom
tom.hosking@ed.ac.uk hao.tang@ed.ac.uk mlap@inf.ed.ac.uk

Abstract

We propose a method for unsupervised abstractive opinion summarization, that combines the attributability and scalability of extractive approaches with the coherence and fluency of Large Language Models (LLMs). Our method, HIRO, learns an index structure that maps sentences to a path through a semantically organized discrete hierarchy. At inference time, we populate the index and use it to identify and retrieve clusters of sentences containing popular opinions from input reviews. Then, we use a pretrained LLM to generate a readable summary that is grounded in these extracted evidential clusters. The modularity of our approach allows us to evaluate its efficacy at each stage. We show that HIRO learns an encoding space that is more semantically structured than prior work, and generates summaries that are more representative of the opinions in the input reviews. Human evaluation confirms that HIRO generates significantly more coherent, detailed, and accurate summaries.

1 Introduction

Online review websites are a useful resource when choosing which hotel to visit or which product to buy, but it is impractical for a user to read hundreds of reviews. Automatic opinion summarization aims to aggregate a large and diverse set of customer reviews about a particular *entity* into a single, easy to read summary. A good summary should accurately reflect the balance of opinions in the input reviews, highlighting the most common or *popular* opinions, while omitting unnecessary details. A useful summary should also help *compare* between competing options, and include points that differentiate the current entity from others.

Early work on opinion summarization extracted reviewers' sentiment about specific features (Hu and Liu, 2004) or identified salient sentences based on centrality (Erkan and Radev, 2004), while more recent methods have proposed *extractive* models that use learned feature spaces (Angelidis et al., 2021; Basu Roy Chowdhury et al., 2022). Prior work on *abstractive* opinion summarization has almost exclusively either required costly supervision (Bražinskas et al., 2021; Cattan et al., 2023) or has assumed that the number of input reviews is limited (Coavoux et al., 2019; Bražinskas et al., 2020; Amplayo et al., 2021a,b; Iso et al., 2021). This defeats the point of a summary: a user could feasibly read 8 reviews in a reasonable period of time. A good summarization system should be *scalable*, since popular products online may receive thousands of reviews. It should also be *attributable*, offering some evidence to justify its output. Paraphrasing Rashkin et al. (2023), we say that a statement s is attributable to some evidence E, if a generic reader would agree that 'According to E, s is true'. Finally, it should generate summaries that are *coherent* and *faithful* to the input reviews.

Large Language Models (LLMs) have been shown to generate highly fluent summaries in the news domain (Bhaskar et al., 2023) but are a flawed solution because current instruction-tuned models are not attributable, and because their context windows limit the number of reviews they are able to base their summaries on. Models with long context windows have been proposed (Beltagy et al., 2020; Gu et al., 2022) but these are not currently instruction-tuned, and it has been shown that LLMs are biased toward information at the start and end of the input (Liu et al., 2023).

Our approach, **H**ierarchical **I**ndexing for **R**etrieval-Augmented **O**pinion Summarization (HIRO), identifies informative sentences using hierarchical indexing and then passes the selected sentences as input to a LLM, similar to retrieval-augmented generation (RAG, Lewis et al., 2020). By separating the steps of content selection and generation, we can combine the attributability and scalability of the discrete

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1533–1555, 2024. https://doi.org/10.1162/tacl_a_00703
Action Editor: Fei Liu. Submission batch: 2/2024; Revision batch: 7/2024; Published 11/2024.

representation with the strong generative abilities of LLMs, leading both to a higher quality index and to more informative and coherent output summaries.

HIRO consists of three modules, allowing for increased control, flexibility and interpretability. The **Hierarchical Indexer** is an encoder that maps sentences from reviews to paths through a hierarchical discrete latent space. The **Retriever** uses the index to identify clusters of sentences for each entity that contain popular and informative opinions. These sentence clusters are passed to a **Generator**, a pretrained LLM, that generates coherent summaries that are grounded in the retrieved sentences.

Our contributions are as follows:

- We propose a method for learning an encoder that maps sentences to a path through a semantically structured discrete hierarchy.
- We show how to exploit this discrete hierarchy at inference time to identify clusters of related and prevalent sentences from input reviews.
- We introduce an automatic metric that measures whether generated summaries reflect the input reviews, while penalizing overly generic statements.
- Through extensive experiments on two English datasets from different product domains, we demonstrate that passing these retrieved sentences in a zero-shot manner to a pretrained LLM generates summaries that better reflect the distribution of opinions within the input reviews. Human evaluation shows that summaries generated by HIRO are significantly more coherent and accurate than prior work, and are preferred by annotators.

Our code and dataset splits are available at `https://github.com/tomhosking/hiro`.

2 Related Work

Opinion Summarization Prior approaches to generating summaries of reviews broadly fall into two categories. *Extractive* methods generate summaries by selecting representative sentences from input reviews to use as the summary (Erkan and Radev, 2004; Angelidis et al., 2021; Basu Roy Chowdhury et al., 2022). These types of approach are scalable and inherently attributable, but result in summaries that are overly detailed and not coherent. *Abstractive* methods 'read' input reviews and generate novel language to use as the summary (Bražinskas et al., 2020; Iso et al., 2021). The resulting summaries are more fluent and coherent, but most prior abstractive methods are only able to consider a limited number of input reviews, whether because of combinatoric complexity or a maximum input length.

Hosking et al. (2023b) propose a hybrid method that represents sentences from reviews as paths through a learned discrete hierarchy, then generates output sentences based on frequent paths through this hierarchy. While their method is abstractive as well as attributable and scalable, the highly compressed bottleneck leads to frequent hallucination and generic output with poor coherence.

Louis and Maynez (2023) use a Natural Language Inference (NLI) model to construct 'silver standard' summaries to use as training data for fine-tuning a pretrained language model (T5, Raffel et al., 2020). However, their approach is computationally very expensive, calculating over 1B pairwise entailment relations between sentences in the training data, before fine-tuning a LLM. By contrast, HIRO uses a lightweight indexing encoder, combined with an off-the-shelf LLM that is prompted in a zero-shot manner.

Other work has used specialized training data to train supervised models (e.g., Zhao et al., 2022; Cattan et al., 2023); however, such data is expensive to collect and is not generally available in every language, domain, or setting.

Structured Encodings and Indexes Prior work has investigated how to learn representations of data with richer structure and interpretability. Vendrov et al. (2016) propose learning an embedding space where the ordering of two pairs of samples could be inferred from their relative positions, but their method requires supervision of the correct ordering. Opper et al. (2023) describe a method for learning embeddings that explicitly include structural information, but they focus on representing structures, rather than on learning an ordering within the embedding space itself. Li et al. (2023) learn a tree-based index for passage retrieval concurrently with the dense embedding space, showing that this leads to improved

retrieval performance. However, they focus on retrieving passages relevant to a query, rather than aggregating opinions. Sarthi et al. (2024) use a pretrained embedding model to cluster related sentences then generate a summary using an LLM, and repeat this process iteratively to construct a tree-structured index over inputs. However, this process is expensive and must be repeated for each new set of inputs.

Content Selection The idea of first selecting relevant parts of an input before generating output has been well studied (Kedzie et al., 2018; Puduppully et al., 2019; Amplayo et al., 2021b; Narayan et al., 2023, *inter alia*), and has been shown to be very effective when used in conjunction with LLMs in the form of retrieval-augmented generation (RAG, Lewis et al., 2020). Xu et al. (2023) found that retrieval augmentation is beneficial even when using models that can accept long inputs. Wang et al. (2023) show that including an additional filtering or selection step to RAG is better than naively passing all retrieved documents as input.

Evaluation of Summaries Automatic evaluation of generated summaries is extremely challenging. Prior work has shown that ROUGE (Lin, 2004) scores correlate poorly with human assessments of summary quality (Callison-Burch et al., 2006; Tay et al., 2019; Fabbri et al., 2021; Shen and Wan, 2023; Clark et al., 2023; Aharoni et al., 2023). Some datasets are created automatically, with references that are not directly based on the input reviews (Bražinskas et al., 2021). Modern summarization system outputs are now often preferred to human-written references (Goyal et al., 2022; Bhaskar et al., 2023).

SummaC (Laban et al., 2022) is a *reference-free* metric that uses an NLI model to evaluate the degree of support between a summary and the input documents, but it overly rewards trivial statements; using the obvious statement ‘‘The hotel was a building’’ as a summary for every entity achieves a near-perfect SummaC score of 99.9% on SPACE, a dataset of hotel reviews (Angelidis et al., 2021).

Malon (2023) proposes a metric that uses a NLI model to evaluate *prevalence*, i.e., how many input reviews contain supporting evidence for each sentence in a summary, and explicitly penalizes trivial or redundant output. However, we found it has a similar failure mode to SummaC, with the statement ‘‘The rooms are clean and comfortable’’ achieving a prevalence score of 72% on SPACE. We propose a modification to prevalence in Section 6.2 that penalizes overly generic summaries.

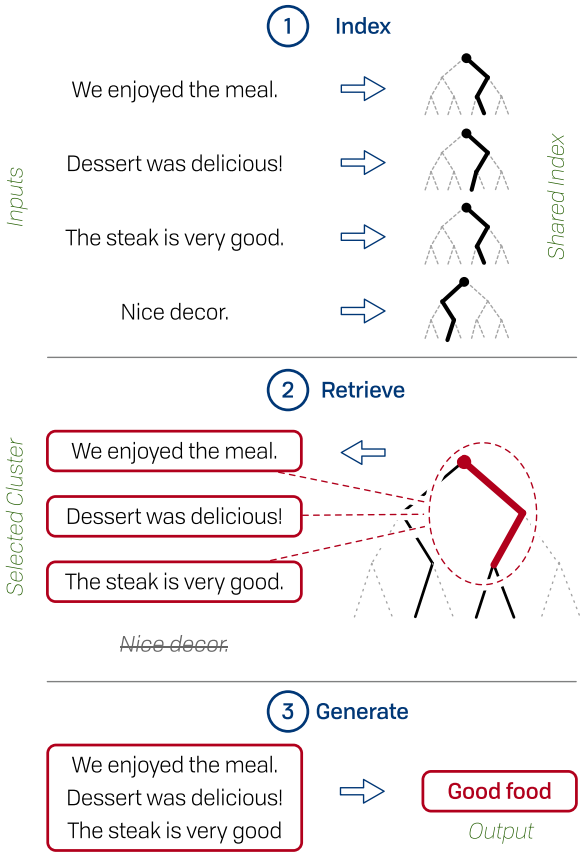

Figure 1: HIRO uses three modules to generate summaries of customer reviews. First, we use our encoder to **index** all sentences from input review into a learned hierarchy. Then we identify paths within this index that occur frequently, and **retrieve** the corresponding clusters of sentences. Finally, we pass these clusters to an LLM to **generate** an output summary.

3 Overview

Let $\mathcal{R}_e$ be a set of reviews about an entity $e \in \mathcal{E}$, where each review $\mathbf{R} \in \mathcal{R}_e$ is composed of a number of sentences $\mathbf{x}$. The goal is to generate a textual summary that includes the most informative opinions from $\mathcal{R}_e$, while abstracting away the details specific to any one review.

HIRO generates summaries following a modular approach, depicted in Figure 1. We learn an index structure that maps each sentence $\mathbf{x}$ from the input reviews to a path $q_{1:D}$ through a discrete hierarchy. We choose a hierarchical representation so that sentences are grouped at a useful level

of abstraction; the upper levels of the hierarchy should partition sentences by topic, while the lowest levels should group together sentences with equivalent meaning.

At inference time, we encode all sentences from the reviews, then identify the paths or *subpaths* $q_{1:d}$ within the index that are particularly *popular*, and retrieve the corresponding sentences. This selection process is query-free, instead relying on properties of the hierarchical index to determine the frequency of different opinions. By indexing sentences hierarchically according to their semantics, we can easily identify opinions or topics that occur frequently by simply counting their occurrence in the index.

Finally, we generate a summary by passing the retrieved clusters of sentences as input to a LLM. This *retrieval-augmented* usage of LLMs allows us to benefit from the fluency and coherence of LLMs, while retaining the attributability and scalability of extractive opinion summarization methods.

We now detail the three modules that constitute HIRO, evaluating each of them in turn to confirm their efficacy compared to previous methods.

4 Learning a Hierarchical Indexing

Our goal is to learn an encoding space where sentences with similar meanings are grouped together. The space should be *discretized* so that frequent opinions can be easily identified by counting the membership of each part of the index, and it should be *hierarchical* so that opinions may be aggregated at an appropriate level of granularity, rather than by details or phrasings specific to a particular review. Finally, the encoding space should be structured semantically, to enable accurate aggregation of opinions; sentences with equivalent meaning should clearly be indexed to the same point in the hierarchy, while sentences that are topically related but not equivalent should be grouped together at a higher level.

We base our encoder on Hercules (Hosking et al., 2023b). Hercules uses an encoder-decoder architecture with a discrete hierarchical bottleneck to generate summaries. It is trained as a denoising autoencoder, and therefore needs to learn a representation that is both compressed enough to enable aggregation, but also expressive enough for the decoder to be able to generate meaningful output. These factors are in direct competition, with the compressed bottleneck leading to output that is generic and contains hallucinations. By contrast, HIRO uses an external LLM as the 'decoder', allowing us to focus on learning a representation that is useful for identifying informative opinions.

4.1 Method

The HIRO encoder module maps a single sentence $\mathbf{x}$ to a path $q_{1:D}$ through a discrete hierarchy, using a technique based on residual vector quantization (Chen et al., 2010; Zeghidour et al., 2022; Hosking et al., 2023b).

First, we use a Transformer encoder followed by attention pooling (Vaswani et al., 2017; Liu and Lapata, 2019) to map a sequence of tokens $\mathbf{x}$ to a single dense embedding $\mathbf{z} \in \mathbb{R}^{\mathbb{D}}$. Then, we decompose $\mathbf{z}$ into a path through a latent discrete hierarchy $q_{1:D}$, where $q_d \in [1, K]$ are discrete 'codes' at each level d. Briefly, we induce a distribution over codes at each level $p(q_d)$, parameterized by a softmax with scores s_d given by the Euclidean distance from learned codebook embeddings to the residual error between the input and the cumulative embedding from all previous levels,

$$s_d(q) = -\left(\left[\mathbf{z} - \sum_{d'=1}^{d-1} \mathbf{C}_{d'}(q_{d'})\right] - \mathbf{C}_d(q)\right)^2, \tag{1}$$

where $\mathbf{C}_d \in \mathbb{R}^{K \times \mathbb{D}}$ is a codebook which maps each discrete code to a continuous embedding $\mathbf{C}_d(q_d) \in \mathbb{R}^{\mathbb{D}}$. During training, we use the Gumbel reparameterization (Jang et al., 2017; Maddison et al., 2017; Sønderby et al., 2017) to sample from the distribution $p(q_d)$. During inference, we set $q_d = \arg\max s_d$.

Since our goal is to learn a representation where semantically similar sentences are grouped together, we use a training objective that explicitly induces this arrangement in encoding space. We train the encoder with a contrastive learning objective, bringing representations of semantically similar sentences (i.e., positive pairs) together, and pushing dissimilar ones apart.

For each sentence in the training data, we construct positive pairs of semantically related sentences $\mathbf{x}, \mathbf{x}_+$ as follows: given a random 'query' sentence $\mathbf{x}$ from the training data, we

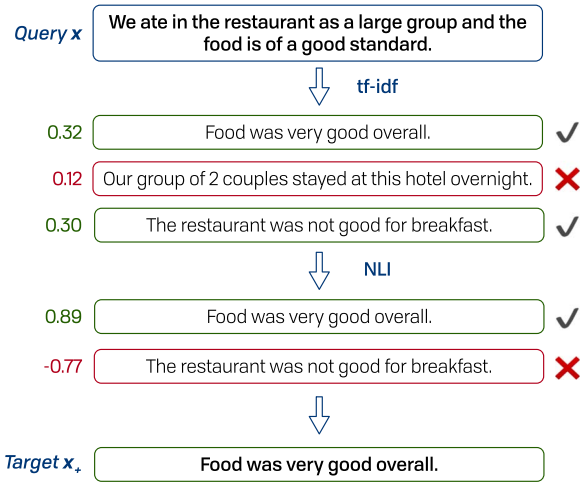

Figure 2: An example of the process for constructing the positive pairs used to train our model. Given a query sentence, we first use tf-idf to identify possible candidates from the training data, keeping only those sentences with similarity over a specified threshold. Then we check for entailment using an NLI model, and use any sentences labelled as 'entailed' as positive targets.

identify possible candidate 'targets' based on tf-idf similarity; then, we check whether each candidate is entailed by the query using an NLI model (DeBERTa v3, trained on Debiased NLI; He et al., 2021; Wu et al., 2022), and use the sentences which are labelled as entailed as positive targets $\mathbf{x}_+$. An example is shown in Figure 2. We do not use any 'hard' negatives; instead, during training we calculate the pairwise tf-idf similarity between all samples in a batch, and include only those samples with similarity below a threshold as negatives $\mathbf{X}_-$. We set this threshold to 0.3 based on validation set performance. This prevents 'false negatives' being wrongly forced apart, and allows sentences that are topically related but not strictly equivalent to remain in the same high-level grouping within the index.

We found that it was crucial to include sufficient exploration in the process of constructing positive pairs. The candidate sentences retrieved using tf-idf should be sufficiently similar that we are likely to find ones that are entailed, but not so similar that they only have minor lexical differences.

We use a modified version of the InfoNCE training objective (van den Oord et al., 2018) designed to bring the representations of a query $\mathbf{x}$ closer to those of a positive target $\mathbf{x}_+$, while pushing them apart from negative samples $\mathbf{X}_-$,

$$\mathcal{L} = -\rho(\mathbf{x}, \mathbf{x}_+) \log f, \tag{2}$$

$$f = \frac{\exp\left(s(\mathbf{x}, \mathbf{x}_+)\right)}{\exp\left(s(\mathbf{x}, \mathbf{x}_+)\right) + \frac{\omega}{|\mathbf{X}_-|} \sum_{\mathbf{x}_- \in \mathbf{X}_-} \exp\left(s(\mathbf{x}, \mathbf{x}_-)\right)},$$

where $\rho(\mathbf{x}, \mathbf{x}_+)$ is the tf-idf similarity between $\mathbf{x}$ and $\mathbf{x}_+$ that weights the *confidence* of the positive pairs, inspired by MarginMSE (Hofstätter et al., 2021), and ω is a constant that controls the strength of the negative examples. The similarity function $s(\cdot, \cdot)$ is given by the mean dot product between the embedding of all subpaths $q_{1:d}$ for $d \leq D$,

$$s(\mathbf{x}, \mathbf{x}') = \frac{1}{D} \sum_{d=1}^{D} \max\left(\mathbf{C}(q_{1:d})^T \mathbf{C}(q'_{1:d}), 0\right), \tag{3}$$

where $\mathbf{C}(q_{1:d}) = \sum_d \mathbf{C}_d(q_d)$ is the full embedding of path $q_{1:d}$. Intuitively, this brings together the representations of the positive pairs at each level in the hierarchy, while penalizing any overlap with the representations of negative examples.

The similarity is floored at zero, to prevent the model from being able to 'game' the loss by pushing negative examples further and further apart. Although the positive pairs are selected based on a directional entailment label, we do not exploit this directionality in our training objective.

We employ the techniques to induce a hierarchical encoding space proposed by Hosking et al. (2023b), including depth dropout, initialization decay, and norm loss. We additionally include the entropy of the distribution over codes, $-\sum_{d,q_d} \log\left(p(q_d)\right)$, as an additional term in the objective, to encourage exploration of the latent space during training.

4.2 Evaluation

We now evaluate whether the combination of discrete hierarchical encoding and contrastive objective leads to a more semantically distributed representation than previous methods.

Experimental Setup We experiment using SPACE (Angelidis et al., 2021), which consists of hotel reviews from TripAdvisor with 100 reviews per entity, as well as reference summaries created by annotators. We encode all the review sentences from the SPACE test set, then calculate

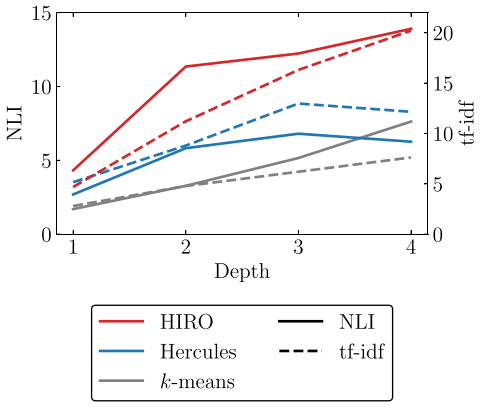

Figure 3: Cluster quality by depth, as measured by the difference between cluster purity and colocation for the SPACE test set, according to NLI (solid line) and tf-idf (dashed line) similarity measures. HIRO learns a higher quality encoding space than comparison methods.

both the *purity* (the mean intra-cluster similarity) and *colocation* (the mean inter-cluster similarity) for the clusters of sentences assigned to each subpath $q_{1:d}$ for $d \leq 4$. Finally, we take the difference between the purity and colocation as an overall measure of the quality of the representation space.

We compare to Hercules (Hosking et al., 2023b) and a baseline using recursive k-means over embeddings from a pretrained embeddings model (MiniLM, Reimers and Gurevych, 2019). We apply k-means to the embeddings, then calculate the residual errors between cluster centroids and embeddings, apply k-means to those errors, and repeat. All methods use the same number of clusters at each level ($K = 12$).

Model Configuration We use a 6 layer Transformer, with token embeddings initialized from BERT base (Devlin et al., 2019).[1] We set the codebook size $K = 12$, with the number of levels $D = 12$, based on validation set performance. Other hyperparameters are given in Appendix B.

HIRO learns a higher quality encoding space Figure 3 shows the overall quality (using both NLI and tf-idf measures of similarity), indicating that HIRO learns a more semantically distributed space at all levels of depth than comparison approaches, according to both similarity measures. The separation between clusters increases with depth, confirming that the encoder learns a semantic hierarchy. We believe these results indicate that our method could potentially be used for more general purpose document retrieval (similar to Li et al., 2023), which is beyond the scope of this paper.

[1]We experimented with using BERT as the encoder but found no significant improvement, since the discrete encoding is the main bottleneck in the model.

5 Retrieving Popular Opinions

Recall that a good summary should include opinions that occur repeatedly in the input reviews, as well as opinions that differentiate the current entity from others. In Section 4 we showed how the HIRO encoder maps a single sentence $\mathbf{x}$ to a path $q_{1:D}$. We now exploit the discrete, hierarchical nature of the representation space to index a large number of review sentences, then identify informative sentences to use to generate a summary. We hypothesize that our content selection method leads to clusters that better represent the balance of opinions in the input reviews.

5.1 Method

For each review $\mathbf{R} \in \mathcal{R}_e$ about an entity $e \in \mathcal{E}$, we separately encode each sentence within the review to its path $q_{1:D}$, giving an indexed review $\mathbf{Q}(\mathbf{R})$.

Our content selection method identifies the parts of the hierarchy that are particularly popular for each entity, and extracts the corresponding clusters of sentences. This process is query-free; instead, we assign each subpath in the hierarchy $q_{1:d}$ a score based on its popularity within the indexed input reviews, and 'retrieve' all sentences that were mapped to the k highest-scoring subpaths. Our scoring function is inspired by tf-idf (Jones, 1972), which is designed to measure the importance of a particular term with respect to a set of baseline documents. We define the *term popularity* $\mathrm{tp}(q_{1:d}, e)$ of a path as the fraction of indexed reviews for entity e which contain the subpath $q_{1:d}$,

$$\mathrm{tp}(q_{1:d}, e) = \frac{1}{|\mathcal{R}_e|} \sum_{\mathbf{R} \in \mathcal{R}_e} \mathbb{I}\big(q_{1:d} \in \mathbf{Q}(\mathbf{R})\big), \quad (4)$$

where $\mathbb{I}$ is the indicator function. We define the *inverse baseline popularity* ibp as the reciprocal of the mean term popularity across *all entities* $\mathcal{E}$,

$$\mathrm{ibp}(q_{1:d}) = \left(\frac{\alpha + \sum_{e \in \mathcal{E}} \mathrm{tp}(q_{1:d}, e)}{\alpha + |\mathcal{E}|} \right)^{-1}, \quad (5)$$

where the smoothing constant α allows us to balance between absolute and comparative popularity. The overall score is then

$$\text{score}(q_{1:d}, e) = \text{tp}(q_{1:d}, e) \times \text{ibp}(q_{1:d}). \quad (6)$$

Intuitively, the score represents the relative popularity within the current entity of a path $q_{1:d}$ compared to all entities in the dataset.

Our scoring scheme conveniently accounts for the fact that short paths are more common;[2] short paths will also have a higher baseline popularity, leading to an overall score that is effectively normalized for the depth of the subpath.

After retrieving the clusters of sentences corresponding to the top-k highest scoring subpaths, we filter out sentences with very low lexical similarity to other cluster members, and combine any clusters of sentences with high lexical overlap.

5.2 Evaluation

We evaluate whether our method successfully selects clusters of sentences containing informative opinions, using reviews from SPACE, as in Section 4.2. First, we measure the overlap between retrieved clusters and *oracle clusters* constructed by selecting sentences with high overlap to the reference summaries,[3] using the Adjusted Rand Index (ARI, Hubert and Arabie, 1985). Second, we evaluate the average *prevalence* (Malon, 2023) of sentences in retrieved clusters, which uses a NLI model (ALBERT, trained on VitC; Lan et al., 2020; Schuster et al., 2021) to evaluate how many input reviews support each retrieved sentence.

We compare HIRO to the clusters extracted by Hercules (Hosking et al., 2023b), and the oracle upper bound. As a baseline, we apply k-means to MiniLM embeddings (Reimers and Gurevych, 2019), then extract the 25 sentences whose embeddings are closest to each centroid. For HIRO, we select the top $k = 8$ subpaths for each entity, and set the smoothing constant $\alpha = 6$.

HIRO Selects Higher Prevalence Sentences Table 1 confirms that HIRO retrieves clusters that more closely match the oracle clusters, and contain opinions that are more prevalent in the input reviews compared to prior work. The oracle ARI score is less than 1 because some sentences appear multiple times in different clusters.

System	Prevalence	ARI
k-means	38.1	0.59
Hercules	32.3	0.50
HIRO	**46.5**	**0.69**
(Oracle)	48.1	0.73

Table 1: Evaluation of our cluster selection method, compared to a range of baseline approaches. HIRO selects clusters of sentences that more closely match the references and contain more prevalent opinions.

6 Generating Coherent Summaries

Given the retrieved clusters of sentences for each entity, we want to generate a coherent and fluent textual summary. LLMs are well suited to this constrained rewriting task, and we leverage the zero-shot abilities of instruction-tuned models to map clusters of sentences to a readable summary.

6.1 Method

We generate a summary from the retrieved clusters in three ways, with varying trade-offs between coherence and attributability.

HIRO$_{\text{ext}}$ We generate *extractive* summaries by selecting the centroid of each cluster; we compute all pairwise ROUGE-2 scores between sentences in each cluster, and choose the sentence with highest average similarity to other cluster members. This approach is inherently attributable, since each summary sentence is extracted verbatim from a review.

HIRO$_{\text{sent}}$ We generate summaries one sentence at a time by passing the contents of a single cluster as input to an instruction-tuned LLM with a simple prompt that requests a single sentence that summarizes the main points. This leads to more fluent output that is likely to be attributable, since each output sentence has an associated cluster of evidential sentences used to generate it.

HIRO$_{\text{doc}}$ We generate summaries as a single document, by passing the sentences from all retrieved clusters for an entity to the LLM in one go. This gives summaries that are more coherent and less redundant, since the LLM has control over

[2] The presence of a path $q_{1:D}$ for a review also implies the presence of all subpaths $q_{1:d}, d < D$.

[3] SPACE includes multiple reference summaries for each entity; we randomly select one when determining the oracle clusters.

the whole summary. However, it is not possible to identify which cluster was used to generate each part of the summary, and therefore more difficult to determine the attributability of the output.

The ideal balance between coherence and the granularity of associated evidence is likely to vary by application.

Experimental Setup For the variants that require an LLM, we use Mistral 7B Instruct v0.2 to generate summaries from retrieved clusters. Mistral 7B was chosen based on its qualitative performance during model development, but we compare using alternative models in Section 6.4. The LLM is queried in a zero-shot manner, and the prompts used are given in Appendix C. We sample with a temperature of 0.7, and report the mean and standard deviation scores based on 3 samples.

6.2 Automatic Evaluation

Human evaluation is the gold standard (Section 6.3), but automatic metrics remain useful for model development. ROUGE scores are no longer reliable (Section 2), but we nonetheless report them for consistency with prior work. Malon (2023) propose a *prevalence* metric, that uses an NLI model to determine how many input reviews contain supporting evidence for each sentence in the summary, but this suffers from a failure mode that overly rewards generic statements. A good summary should include common opinions, but should also help a user to differentiate between multiple entities.

To counteract this pathology, we propose a modified version of prevalence, that explicitly penalizes generic summaries. First, we define the *genericness* of a summary as the average number of summaries from *other entities* that support each sentence in a given summary, as measured by an NLI model (ALBERT, trained on VitC; Lan et al., 2020; Schuster et al., 2021). Then, we define the Specificity Adjusted Prevalence score (SAP) as

$$\texttt{SAP} = \texttt{prevalence} - \alpha\,\texttt{genericness}, \tag{7}$$

where α is a constant that controls the balance between absolute prevalence and specificity. In practice, the ideal summary is unlikely to be entirely unique and a user may want to allow some degree of overlap between generated summaries. We report the prevalence and genericness, as well as the combined SAP with $\alpha = 0.5$.

Datasets We evaluate summary generation using SPACE (Section 4.2), which includes multiple reference summaries created by human annotators for each entity. We also include **AmaSum** (Bražinskas et al., 2021), to evaluate summary generation on a wide range of categories of Amazon products, with an average of 560 reviews per entity. The reference summaries were collected from professional review websites, and therefore are not necessarily grounded in the input reviews. We use the same splits, based on four product categories, as released by Hosking et al. (2023b). Further dataset statistics are given in Appendix D.

Comparison Systems We select a **random review** from the inputs as a lower bound. We include an extractive **oracle** as an upper bound, by selecting the input sentence with highest ROUGE-2 similarity to each reference sentence.[4] For a **k-means** baseline, we run k-means on MiniLM sentence embeddings (Reimers and Gurevych, 2019), then extract the nearest sentence to the cluster centroids. We set $k = 8$ to match the average sentence length of the reference summaries. **Lexrank** (Erkan and Radev, 2004) is an unsupervised extractive method using graph-based centrality scoring of sentences. **QT** (Angelidis et al., 2021) uses vector quantization to map sentences to a discrete encoding space, then generates extractive summaries by selecting representative sentences from clusters. **SemAE** (Basu Roy Chowdhury et al., 2022) is an extractive method that extends QT, relaxing the discretization and encoding sentences as mixtures of learned embeddings. **CopyCat** (Bražinskas et al., 2020) is an abstractive approach that models sentences as observations of latent variables representing entity opinions, trained in a 'leave one out' manner. **BiMeanVAE** and **COOP** (Iso et al., 2021) are abstractive methods that encode full reviews as continuous latent vectors using an autoencoder, and take the average (BiMeanVAE) or an optimized combination (COOP) of review encodings. We compare to a recent open weight instruction-tuned **LLM**, specifically Mistral Instruct v0.2 7B (Jiang et al., 2023). Since no training data is available, the LLM was prompted zero-shot

[4] When multiple references are available, we select one at random.

	System	SPACE R-2 ↑	SPACE R-L ↑	SPACE Prev. ↑	SPACE Gen. ↓	SPACE SAP ↑	AmaSum (4 domains) R-2 ↑	AmaSum (4 domains) R-L ↑	AmaSum (4 domains) Prev. ↑	AmaSum (4 domains) Gen. ↓	AmaSum (4 domains) SAP ↑
Extractive	Rand. Review	6.2	17.1	18.0	12.5	11.8	1.0	9.5	16.3	8.0	12.3
	k-means	9.5	19.8	27.9	25.0	15.4	2.3	12.0	14.9	11.4	9.2
	LexRank	5.9	16.4	18.2	**4.4**	16.0	2.7	12.2	9.0	**3.0**	7.5
	QT	10.3	21.5	24.9	23.3	13.3	1.5	11.4	10.9	7.3	7.3
	SemAE	11.1	23.5	29.2	17.1	20.6	1.6	11.3	8.7	4.1	6.7
	$\text{Hercules}_{\text{ext}}$	**13.2**	**24.4**	30.2	25.2	17.6	**3.0**	12.5	9.5	6.7	6.2
	HIRO_{ext}	11.7	22.1	**36.3**	20.5	**26.1**	2.7	**12.6**	**19.4**	9.5	**14.6**
Abstractive	CopyCat	12.1	22.9	**48.3**	70.9	12.9	1.5	11.2	15.8	21.0	5.3
	BiMeanVAE	13.7	27.1	45.0	61.4	14.2	2.0	12.5	14.7	24.1	2.7
	COOP	14.2	**27.2**	46.1	63.2	14.5	2.8	14.1	**18.8**	30.3	3.7
	$\text{Hercules}_{\text{abs}}$	**14.8**	**27.2**	32.2	36.1	14.1	2.0	11.8	8.5	9.2	3.9
	Zero-shot Mistral 7B	5.3±0.1	19.6±0.4	41.3±1.3	34.3±3.3	24.2±0.8	1.9±0.0	12.6±0.0	17.3±0.2	17.6±0.4	8.5±0.2
	$\text{HIRO}_{\text{sent}}$+Mistral 7B	4.5±0.1	18.2±0.0	36.2±0.8	**20.1**±0.5	26.2±0.9	3.5±0.0	14.1±0.1	14.6±0.3	**6.9**±0.1	**11.2**±0.3
	HIRO_{doc}+Mistral 7B	7.0±0.2	20.5±0.3	44.0±3.0	28.8±2.1	**29.6**±2.1	**4.0**±0.0	**15.1**±0.1	15.3±0.1	9.4±0.3	10.6±0.1
	(References)	–	–	44.3	50.2	19.2	–	–	9.3	7.0	5.8
	(Oracle)	45.0	53.3	41.0	38.5	21.7	14.4	26.0	12.3	9.0	7.8

Table 2: Results for automatic evaluation of summary generation on the test splits. R-2 and R-L represent ROUGE-2/L F1 scores. Prev. refers to Prevalence, Gen. refers to Genericness, and SAP refers to Specificity-Adjusted Prevalence. ROUGE scores are no longer considered reliable (Callison-Burch et al., 2006; Tay et al., 2019; Fabbri et al., 2021), so we consider SAP to be our primary metric. The best scores for extractive and abstractive systems are shown in bold. Results for systems involving LLMs are based on 3 samples, with the mean and standard deviation shown. HIRO generates summaries with the best balance between prevalent opinions and specificity.

as per Appendix C, and sampled with temperature 0.7. We report the mean and standard deviation scores based on 3 samples.

Most of the abstractive methods are not scalable and have upper limits on the number of input reviews. CopyCat and Mistral 7B have a maximum input sequence length, while COOP exhaustively searches over combinations of input reviews. We use 8 randomly selected reviews as input to CopyCat, COOP, and Mistral 7B.

HIRO Gives the Best Balance Between Prevalence and Specificity The results in Table 2 show that HIRO achieves the highest SAP scores across both datasets, indicating that it generates summaries with the best balance between absolute prevalence and specificity. While CopyCat and COOP achieve the highest prevalence scores on SPACE and AmaSum respectively, they also display some of the highest genericness; qualitatively, the outputs are very similar to each other, with few specific details. We give example output in Tables 10 and 11.

References Are not the Upper Bound While the oracle summaries score highly in terms of specificity-adjusted prevalence, some systems (including HIRO) outperform them. This indicates the difficulty with constructing reference summaries for entities with large numbers of reviews; it is not feasible for human annotators to reliably summarize such a large amount of information.

HIRO Is More Faithful to Selected Evidence To evaluate how faithful the generated summaries are to the retrieved sentence clusters or *evidence sets*, we use an NLI model to determine how many sentences in each cluster either entail or are entailed by the corresponding sentence in the output summary, and take the mean. Considering both forward and backward entailment in this manner accounts for the different levels of granularity between the inputs and summary (Zhang et al., 2024); input reviews are likely to be more specific than summary sentences, but concise summary sentences are likely to contain multiple assertions, e.g., ‘‘The food was good *and* the rooms were clean’’. HIRO_{doc} does not align the evidence sets with each generated sentence, so we calculate the maximum support score over all sets for each summary sentence. Most abstractive systems are

System	% Partial	% Majority
$\text{Hercules}_{\text{abs}}$	85.4	27.6
$\text{HIRO}_{\text{sent}}$	81.6	21.6
HIRO_{doc}	**91.8**	**28.4**

Table 3: Results for automatic evaluation of the evidence supplied by attributable systems, showing the percentage of summary sentences that have support from at least one sentence in the evidence set (partial support) and from at least half the sentences in the evidence (majority support). HIRO generates summaries that have strong partial support from the associated evidence sets, with improved majority support compared to Hercules.

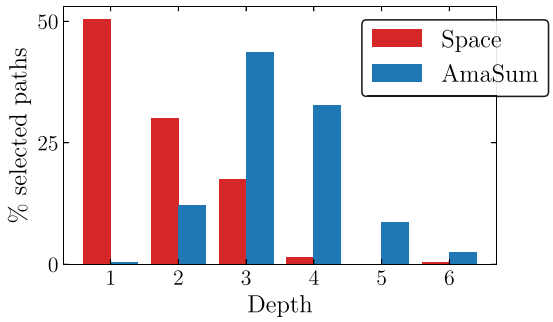

Figure 4: Distribution of selected subpaths by depth. A large fraction of the extracted clusters come from paths deeper than the top level.

not attributable, and so we only compare with Hercules. Table 3 shows the percentage of summary sentences that have support from at least *one* sentence in the evidence (partial support) and from at least *half* the sentences in the evidence (majority support). HIRO_{doc} generates summaries that have strong partial support from the associated evidence sets, with improved majority support compared to Hercules despite also being significantly more detailed.

HIRO Makes Use of the Hierarchy We confirm that HIRO exploits the hierarchical nature of the representation space. Figure 4 shows the distribution of selected subpath depths for both datasets, indicating that HIRO takes advantage of the hierarchy and selects clusters deeper than the top level. This is particularly true for AmaSum, where there is a wider range of product types, causing the selection process to search deeper in the tree.

6.3 Human Evaluation

We conduct a human evaluation to verify that HIRO generates summaries that are coherent and accurately reflect the opinions in the input reviews. We recruit crowdworkers through Prolific, selected to be L1 English speakers from the US or UK with a minimum of 100 previously approved studies, compensated above the UK living wage at 12GBP/hr. Participants were allowed to rate at most 5 samples. We show annotators a set of 50 reviews (chosen based on pilot studies to balance annotator load with reliability), followed by two generated summaries. We solicit pairwise preferences (Louviere and Woodworth, 1990) along three dimensions, as well as an overall preference:

- **Accuracy** — Which summary accurately reflects the balance of opinion in the reviews?
- **Detail** — Which summary includes more specific details?
- **Coherence** — Which summary is easy to read and avoids contradictions?
- **Overall** — Which summary do you think is better, overall?

The full instructions are reproduced in Appendix A. Ties (i.e., 'no difference') were allowed. We gather annotations for 10 entities each from the SPACE and AmaSum test sets, with 3 annotations for each pairwise combination of system outputs, leading to a total of 900 pairwise ratings. The study was approved by the ethics committee, ref. #491139.

We compare HIRO_{doc} to the top performing extractive and abstractive systems from Table 2, SemAE and $\text{Hercules}_{\text{abs}}$. HIRO_{doc} uses Mistral 7B as the generator, so we also compare to Mistral 7B *without* HIRO (i.e., prompted directly with reviews). Finally, we include a random review as a lower bound, and the references as an upper bound.

Humans Prefer Summaries Generated by HIRO The results in Figure 5 show that HIRO_{doc} produces summaries that outperform comparison systems across all dimensions, producing summaries that coherently and accurately represent the opinions in the input reviews. Differences between HIRO_{doc} and other systems are significant in all cases (using a one-way ANOVA with post-hoc Tukey HSD test, $p < 0.05$), except for coherence versus Mistral 7B. Both Mistral 7B

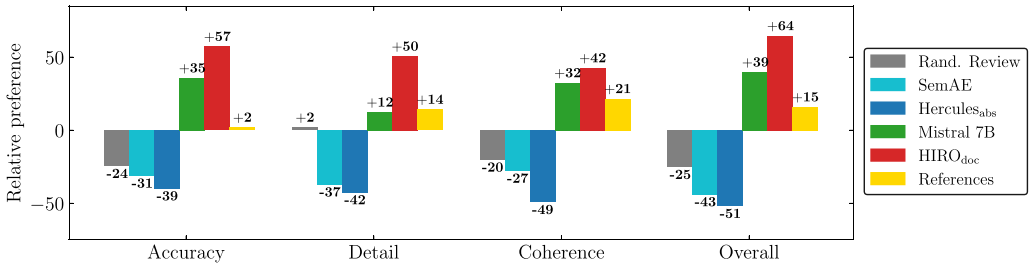

Figure 5: Overall results of our human evaluation based on the test splits of SPACE and AmaSum. Crowdworkers were asked for pairwise preferences between generated summaries in terms of their Accuracy, Detail, and Coherence, & Fluency, as well as an overall preference. HIRO generates more coherent and detailed summaries that better represent the opinions in the input reviews than comparison systems, and are preferred by human annotators.

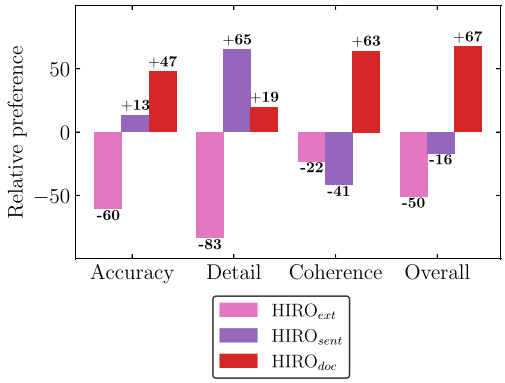

Figure 6: Results of a human evaluation comparing the three variants of HIRO: extractive, sentence-wise and document. Overall, annotators prefer the coherence of the document approach, but the sentence-wise variant generates more detailed summaries that are also more attributable. The preferred tradeoff between coherence and attribution will vary depending on the application.

and HIRO$_{doc}$ outperform the reference summaries, supporting findings from prior work (Bhaskar et al., 2023; Hosking et al., 2023a).

Comparing HIRO Variants We run an additional human evaluation to evaluate the tradeoff between quality and attribution between the three HIRO variants, collecting annotations for 5 entities from each dataset from 3 annotators each. Figure 6 shows that annotators prefer the output of HIRO$_{doc}$, although the sentence-wise variant generates more detailed summaries. The preferred tradeoff between attribution and coherence will likely vary depending on the application.

6.4 Analysis

Ablations Table 4 compares HIRO to zero-shot prompting, for a range of different LLM sizes from the Llama 2 Chat family of LLMs (Touvron et al., 2023). The same prompt template and hyperparameters were used as for Mistral 7B. Increased parameter count does lead to improved SAP scores for the zero-shot approach on SPACE, but the largest 70B model scores worse on AmaSum than the 7B and 13B models. For all choices of LLM, HIRO leads to summaries with a better balance between prevalence and specificity than zero-shot prompting; HIRO is likely to lead to improvements for any choice of instruction-tuned LLM.

We also compare using HIRO to identify clusters against a k-means baseline, and Hercules (Hosking et al., 2023b). For both datasets, using clusters selected with HIRO leads to summaries that are much less generic, while remaining comparatively prevalent. This confirms the results in Section 5.2, indicating that HIRO selects more informative clusters of sentences than the comparison methods.

Qualitative Analysis Table 5 shows output from Mistral 7B and HIRO as well as a reference summary, for The Grand Hotel Maryland from SPACE. While all generated summaries are fluent and convincing, Mistral 7B makes reference to the opinion of a single user, which should not appear in a summary. It also describes positive praise about a member of staff, but a manual analysis shows that only 3 out of 5 reviews mentioning that individual are positive; the true sentiment is much more mixed than the summary indicates. These extrapolations highlight the problem with models that can only accept a limited number of reviews as input.

The examples of selected subpaths, sentence clusters and corresponding HIRO outputs in

	Clusters	LLM	SPACE R-2 ↑	SPACE R-L ↑	SPACE Prev. ↑	SPACE Gen. ↓	SPACE SAP ↑	AmaSum R-2 ↑	AmaSum R-L ↑	AmaSum Prev. ↑	AmaSum Gen. ↓	AmaSum SAP ↑
	Zero-shot	Llama 2 7B	4.4	17.6	32.6	25.7	19.7	1.5	11.5	**17.7**	17.6	8.9
Llama 2 7B	$\text{HIRO}_{\text{sent}}$	Llama 2 7B	5.2	16.9	36.2	**20.0**	26.2	3.6	14.0	14.6	**6.3**	**11.5**
	HIRO_{doc}	Llama 2 7B	**6.9**	**19.9**	**48.5**	29.9	**33.5**	**4.0**	**15.2**	16.0	12.1	10.0
	Zero-shot	Llama 2 13B	6.3	19.1	36.7	28.1	22.6	2.3	13.1	18.1	14.3	10.9
Llama 2 13B	$\text{HIRO}_{\text{sent}}$	Llama 2 13B	6.6	18.8	35.0	**23.7**	23.1	3.8	14.3	13.5	**6.2**	10.4
	HIRO_{doc}	Llama 2 13B	**8.5**	**21.7**	**46.2**	27.0	**32.7**	**4.3**	**15.7**	**19.0**	9.9	**14.1**
	Zero-shot	Llama 2 70B	5.6	19.4	**48.9**	37.2	30.3	2.1	12.5	16.8	22.4	5.6
Llama 2 70B	$\text{HIRO}_{\text{sent}}$	Llama 2 70B	5.8	18.5	41.1	**19.6**	31.3	3.6	14.5	15.1	**3.7**	**13.2**
	HIRO_{doc}	Llama 2 70B	**8.3**	**21.3**	48.5	30.0	**33.5**	**4.4**	**16.0**	**17.8**	9.9	12.9
	k-means	Mistral 7B	**4.5**	17.1	30.1	30.8	14.7	3.2	13.3	12.9	12.4	6.7
Sent-wise	Hercules	Mistral 7B	3.9	17.0	27.4	25.8	14.5	3.3	11.8	8.5	8.7	4.1
	$\text{HIRO}_{\text{sent}}$	Mistral 7B	**4.5**	**18.2**	**36.4**	**20.2**	**26.3**	**3.5**	**14.1**	**14.6**	**6.9**	**11.2**
	k-means	Mistral 7B	6.4	20.5	40.2	36.3	22.0	3.7	14.5	**16.2**	13.9	9.2
Doc-wise	Hercules	Mistral 7B	**7.6**	**21.0**	42.9	39.2	23.3	3.8	14.7	13.8	11.6	8.0
	HIRO_{doc}	Mistral 7B	7.0	20.5	**44.0**	**28.8**	**29.6**	**4.0**	**15.1**	15.3	**9.4**	**10.6**

Table 4: Automatic evaluations comparing HIRO to zero-shot summarization, using a range of different LLMs, for the test splits. We show the mean scores based on 3 samples, and best scores within each comparison are bolded. In all cases, HIRO improves on the zero-shot approach. We also compare HIRO to other cluster selection methods, finding that HIRO leads to summaries with a better balance between prevalance and specificity.

System	Output
Reference	The staff were very friendly, courteous and helpful. The hotel and room was clean. The rooms are very spacious with a nice oceanfront view and a gigantic balcony! Pool bar is great and food is good, too. The hotel has a perfect location. It was very nicely located with a 24 hours cafe downstairs, 2 pools, 2 bars and some beach stores to shop in.
Mistral 7B	The Grand Hotel in Ocean City, Maryland, consistently receives positive reviews for its exceptional customer service, cleanliness, and convenient location on the boardwalk. Rooms offer great views, and amenities include a spa, restaurant, lounge, indoor pool, game room, and exercise room. Guests appreciate the family-friendly atmosphere and the ease of having everything they need within the hotel. Some reviewers praise Nina Railean for her helpfulness and accommodating nature. However, one review warns against staying during Memorial Day weekend due to loud parties.
$\text{HIRO}_{\text{sent}}$	All rooms offer excellent ocean views at this hotel. The Grand Hotel & Spa offers direct ocean views and easy beach access. The Grand Hotel & Spa offers clean and spacious rooms. No problems with accommodations or access despite rain and floor issues. The Grand Hotel & Spa offers beautiful views and a comfortable stay with excellent service. The Grand Hotel & Spa exceeded expectations. Rooms need updating peeling wallpaper, large, dark interiors.
HIRO_{doc}	The Grand Hotel & Spa offers stunning ocean views from most rooms, with some having a direct view of the boardwalk and beach. The hotel's design ensures that guests can enjoy the ocean despite room location. The beach is clean, and the location is convenient for easy access. Rooms are described as clean and large, but some note a need for improvement in terms of lighting and room condition. Overall, guests had positive experiences and appreciated the ocean access.

Table 5: Generated summaries for The Grand Hotel Maryland, from SPACE, comparing HIRO to zero-shot Mistral 7B. While all generated summaries are fluent and convincing, Mistral 7B makes reference to the opinion of a single user, and positive sentiment about a member of staff that is not supported by the full set of reviews. These extrapolations highlight the problem with models that can only accept a limited number of reviews as input.

Table 12 demonstrate the difficulty of evaluating attribution. The final cluster contains sentences that are all topically related, indicating that HIRO has learned a successful clustering. While the sentences and corresponding $\text{HIRO}_{\text{sent}}$ output are all broadly negative, it is not straightforward to determine whether the sentence ''Only one mirror in the room'' counts as direct evidence towards

HIRO provides more specific detail about user frustrations and the product itself. Mistral 7B is quite generic.
Mistral 7B mentions that desk staff were unfriendly, but this is not substantiated by the reviews, the majority of which are overwhelmingly positive. It's also a contradiction since earlier it says the staff were friendly.
Only HIRO references the high failure rate.
HIRO seems more accurate in its appraisal of the staff; the reviews were mixed. However Mistral 7B is slightly more informative, particularly in relation to location and proximity to amenities. I feel like HIRO sits on the fence quite a lot, whereas Mistral 7B attempts to summarize better.
I think HIRO is the better written and reflects a broader view on the reviews but References isn't bad it does cover PS4 and gaming which alot of reviews were using it for.
SemAE is incoherent. HIRO does an excellent job of summarising everything, balancing all perspectives.
Can tell HIRO is AI generated
HIRO is by far the better written and has a little more detail and reflects the reviews more than $\text{Hercules}_{\text{abs}}$

Table 6: Comments from annotators for all pairs involving HIRO. Anonymized system labels have been replaced with the system names. The majority of comments are positive towards HIRO, although annotators comment that HIRO summaries may be less natural, or too conservative.

the output ''Rooms were small, loud, and in need of renovation [...]''. This partly explains the relatively low majority support scores in Table 3; some of the selected evidence may be consistent in topic and sentiment, but not directly entail the resulting output.

As well as collecting pairwise preferences in our human evaluation, we allowed annotators to leave qualitative comments. Table 6 shows all non-trivial comments from annotators for pairs including HIRO, with anonymized labels replaced by the true model names. The majority of comments are positive towards HIRO, highlighting improved levels of detail and a better balance of the input reviews. However, some annotators note that HIRO summaries may be less natural or too conservative.

7 Conclusion

We propose HIRO, a modular method for unsupervised opinion summarization that uses a hierarchical index over sentences to select clusters of prevalent opinions. Our approach leverages pretrained Large Language Models to generate coherent and fluent summaries that are attributable and accurately reflect the popularity of opinions in the input reviews. Extensive experiments show that, as well as generating higher quality summaries, HIRO learns a more semantically distributed representation than competitive baselines. While we limit our experiments to opinion summarization, we believe that HIRO could be usefully applied to a wide range of other retrieval-augmented generation tasks.

Acknowledgments

We thank the action editor and anonymous reviewers for their constructive feedback. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh. Lapata acknowledges the support of the UK Engineering and Physical Sciences Research Council (grant EP/W002876/1).

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.220

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. https://doi.org/10.18653/v1/2021.emnlp-main.528

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021b. Unsupervised opinion summarization with content planning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2–9, 2021*, pages 12489–12497. AAAI Press. https://doi.org/10.1609/aaai.v35i14.17481

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. https://doi.org/10.1162/tacl_a_00366

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. https://doi.org/10.18653/v1/2022.acl-long.86

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.591

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. https://doi.org/10.18653/v1/2020.acl-main.461

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. https://doi.org/10.18653/v1/2021.emnlp-main.743

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Arie Cattan, Lilach Eden, Yoav Kantor, and Roy Bar-Haim. 2023. From key points to key point hierarchy: Structured and expressive opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 912–928, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.52

Yongjian Chen, Guan Tao, and Cheng Wang. 2010. Approximate nearest neighbor search by residual vector quantization. *Sensors (Basel, Switzerland)*, 10:11259–73. https://doi.org/10.3390/s101211259

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.584

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. https://doi.org/10.18653/v1/D19-5405

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. https://doi.org/10.18653/v1/N19-1423

G. Erkan and D. R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479. https://doi.org/10.1613/jair.1523

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409. https://doi.org/10.1162/tacl_a_00373

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *arXiv preprint*.

Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2021. Improving efficient neural ranking models with cross-architecture knowledge distillation.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2023a. Human feedback is not gold standard.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023b. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.473

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*, KDD '04, pages 168–177, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1014052.1014073

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218. https://doi.org/10.1007/BF01908075

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. https://doi.org/10.18653/v1/2021.findings-emnlp.328

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with Gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21. https://doi.org/10.1108/eb026526

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1208

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177. https://doi.org/10.1162/tacl_a_00453

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Haitao Li, Qingyao Ai, Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Zheng Liu, and Zhao Cao. 2023. Constructing tree-based index for efficient and effective dense retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 131–140, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3539618.3591651

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.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1500

Annie Louis and Joshua Maynez. 2023. OpineSum: Entailment-based self-training for abstractive opinion summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10774–10790, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.686

Jordan J. Louviere and George G. Woodworth. 1990. Best worst scaling: A model for largest difference judgments [working paper]. *Faculty of Business*.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*. OpenReview.net.

C. Malon. 2023. Automatically evaluating opinion prevalence in opinion summarization. In *The 6th Workshop on e-Commerce and NLP (KDD 2023)*.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996. https://doi.org/10.1162/tacl_a_00583

Mattia Opper, Victor Prokhorov, and Siddharth N. 2023. StrAE: Autoencoding for pre-trained embeddings using explicit structure. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7544–7560, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.469

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *The Thirty-Third*

AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 – February 1, 2019, pages 6908–6915. AAAI Press. https://doi.org/10.1609/aaai.v33i01.33016908

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

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, 49(4):777–840. https://doi.org/10.1162/coli_a_00486

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. https://doi.org/10.18653/v1/D19-1410

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations (ICLR)*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! Robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643. Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.52

Yuchen Shen and Xiaojun Wan. 2023. Opinsummeval: Revisiting automated evaluation for opinion summarization.

Casper Kaae Sønderby, Ben Poole, and Andriy Mnih. 2017. Continuous relaxation training of discrete latent variable image models. In *Beysian DeepLearning workshop, NIPS*, volume 201.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

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.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.190

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. Soundstream: An end-to-end neural audio codec.*IEEE/ACM Transactions on Audio, Speech and Language Processing*, 30:495–507. https://doi.org/10.1109/TASLP.2021.3129994

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

Mengxue Zhao, Yang Yang, Miao Li, Jingang Wang, Wei Wu, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2022. Personalized abstractive opinion tagging. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1066–1076, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3477495.3532037

A Annotation Instructions

Instructions

In this task you will be presented with some reviews of a product/hotel, followed by two summaries produced by different automatic systems. Your task is to rate the system summaries based on the criteria listed below.

First, please skim read through the reviews, to try to get an overall idea of what opinions are mentioned frequently. Then, read the system summaries carefully, and rate each one according to the criteria.

Please read the criteria descriptions and system summaries carefully, and whenever is necessary re-read the summaries or reviews. You might want to use your browser's search function to help find parts of reviews that are relevant.

Criteria

Accuracy – Which system summary accurately reflects the balance of opinion in the input reviews? Statements in the summary should be backed up by multiple reviews.

Detail – Which system summary includes more specific details?

Coherence & Fluency – Which system summary is easy to read and avoids contradictions?

Overall – Which system summary do you think is better, overall?

B Hyperparameters

Parameter	Value
Embedding dim. $\mathbb{D}$	768
Encoder layers	5
Feedforward dim.	2048
Transformer heads	8
Depth D	12
Codebook size K	12
Optimizer	Adam (Kingma and Ba, 2015)
Learning rate	1e–4
Batch size	384
ω	150
α_{init}	0.5
τ_0	1.0
τ_{min}	0.5
γ_{temp}	33333
β_{KL}	0.0025
β_{NL}	0.05
γ_{NL}	1.5
top-k subpaths	8
$\mathrm{tp} - \mathrm{ibp}$ smoothing α	6 (SPACE), 3 (AmaSum)

Table 7: Hyperparameter values used for our experiments.

C LLM Prompts

```
Review:
[...] (x8)
Write a summary in 70 words or
   less:
```

Prompt 1: Baseline LLMs (Mistral 7B and all Llama 2 models)

```
Here is a list of sentences
   taken from reviews of the
   {entity name}:
[...]

In no more than 10 words, write
   a single concise sentence
   that includes the main point:
```

Prompt 2: $\text{HIRO}_{\text{sent}}$

```
Here is a list of sentences
   taken from reviews of the
   entity name:

[...]

In no more than 60 words,
   write a concise summary that
   includes the main points:
```

Prompt 3: HIRO_{doc}

D Dataset Statistics

	SPACE	AmaSum
Entities	8350	7255
Reviews	303,357	533,972
Training pairs $(\mathbf{x}, \mathbf{x}_+)$	1,373,079	2,991,478

Table 8: Statistics for the training datasets.

	SPACE	AmaSum
Entities	25	200
Reviews per entity	100	560.4
Review length (words)	162.6	49.9
Ref. summaries (words)	82.0	80.1

Table 9: Statistics for the evaluation datasets.

E Example Outputs

See Table 10, Table 11, and Table 12.

System	Output
Reference	The staff were very welcoming and were always happy to help you with whatever was needed. The comfortably-sized rooms were very nice and clean, and the bathrooms were exceptionally clean and well equipped. Breakfast was continental and included in the price, offering baguettes, croissants, cereal, yogurts, etc.
Random Review	A real parisienne experience. The hotel is a beautiful building and the location perfect. We walked everywhere and the local area is a delight with many bars and FRIENDLY restaurants. We are an older couple and went to Paris to get engaged - it was all perfect and the hotel's ambience added to the experience. Highly recommended if you want a small, local hotel as opposed to a large modern one. Many thanks - we would return and recommend you to friends. Margaret and David
SemAE	We were able to walk to all the tourists sights, except Montmarte and the Eiffel Tower. The hotel staff was very friendly and helpful during our stay. The staff is very friendly and helpful and the breakfast is a perfect start to the day. The rooms was sufficent in size, the bed and pillows were very comfortable. The ladies in the breakfast area were very friendly and courteous. The bathroom was clean and the shower
CopyCat	We stayed at this hotel for 3 nights. The hotel was very clean and the staff very helpful. The location is perfect for exploring the city and the metro station is a short walk away. We will be back!
COOP	The hotel is in a great location, close to the metro and the metro. The staff were very friendly and helpful. The rooms were clean and comfortable, and the breakfast was very good. The hotel is in a great location, close to the Metro and the metro. The only downside is that the rooms are a little on the small side, but they are very clean and the staff is very friendly and helpful.
Hercules$_{abs}$	The room was clean and comfortable. The staff was very friendly and helpful. walking distance to everything. The room was clean and comfortable. breakfast was good. The hotel is in a great location, just a few minutes walk from the train station. breakfast was fine. The room and bathroom were very clean. The staff spoke English and were very helpful. There is also a small restaurant on the ground floor.
Mistral 7B	The St. Christophe Hotel in Paris offers small, clean, and cozy rooms in a lively neighborhood near the Place Contrascarpe and Rue Mouffetard. The staff are helpful and friendly, and the location is convenient with easy access to metro stations and nearby restaurants. The hotel lacks air conditioning, so it may not be ideal during warm weather. Free Wi-Fi is available in the lobby. Overall, the St. Christophe Hotel is a charming and elegant choice for travelers to Paris.
HIRO$_{ext}$	Nearby, and very close to 2 Metro stations. The hotel is 1 block to the Metro (line 7) between the North and South exits of the ''place Monge'' stop. The bathroom was clean and functional. The staff were friendly and helpful at all times. Breakfast was Plentiful-Cold cereal, great coffee, yogurt and a wonderful basket of Croissants, breads and Pastries. The shower excellent.
HIRO$_{sent}$	The Hotel Saint Christophe is conveniently located near multiple metro stations. The Hotel Saint Christophe is centrally located near the Metro and restaurants in the Latin Quarter. The Hotel Saint Christophe offers clean, functional bathrooms in small rooms. The Hotel Saint Christophe's friendly and helpful staff provided excellent service. Breakfast offers cereal, yogurt, pastries, and hot beverages. Guests praised the Hotel Saint Christophe for its excellent showers.
HIRO$_{doc}$	The Hotel Saint Christophe is centrally located in the Latin Quarter, close to multiple Metro stations and attractions such as Notre Dame and Jardin Luxembourg. Rooms are clean and adequate, though small, and breakfast offers a variety of options. The staff are friendly and helpful, making for a pleasant stay. The hotel is conveniently located near shops, restaurants, and markets. The bathrooms are clean and modern, though some may be smaller than expected for Paris.

Table 10: Examples of generated summaries for The Hotel Saint Christophe from SPACE. HIRO$_{doc}$ generates concise, informative and coherent summaries. Mistral 7B prompted directly with reviews also generates coherent summaries, but is limited to a small sample of reviews which do not necessarily reflect the overall balance of opinions. Previous abstractive systems, CopyCat, COOP and Hercules$_{abs}$, generate generic summaries that lack detail.

System	Output
Reference	These headphones are your best option if you're looking for noise-cancelling earbuds, but they are pricey. Noise-cancelling earbuds. Available in black or white. Stay Hear tips keep earbuds in your ears while maximizing comfort. Impressive sound quality. Carrying case included. These are expensive for earbuds, so you should only buy them if you're going to use them often
Random Review	Excellent. I'm comparing this to the Bose on ear headphones and this cancels noise better. I'm a frequent flyer and like to disappear during flights. This helps me get into another zone whether it's listening to electro-house or watching the latest movies while flying. Highly recommend it!
CopyCat	So far so good. The only issue I have is that the volume is not as loud as my old one, but other than that I'm happy with this purchase.
COOP	These are the best ear buds out of the box. The sound quality is great, the bass is very good, and the noise cancelling feature is amazing. The only downside is that it doesn't have a USB-C port on the back of the device. It's not a big deal, but if you want to use it for work, you can't go wrong with this product.
$Hercules_{abs}$	Love these Earbuds. These are great for listening to music. Noise cancelling works well. Long battery life. These are my favorite Headphones. The sound quality is very clear. The sound quality is great and the noise cancellation is good. Got this for my wife for Christmas. The sound quality is great and the Mic is clear. The Bass is a bit lacking. The ear pads are soft and comfortable. These Headphones are the best ever. The Headphones are very comfortable.
Mistral 7B	These earbuds have received positive reviews from users who travel frequently, find them a game changer, and have even convinced friends to buy them. Some users have purchased them for their children who are also satisfied. The earbuds have received five-star ratings from both the buyers and their sons. One user's wife loves them and uses them when the user snores at night. However, some users have noted that the noise-cancelling feature is not as effective as over-ear headphones.
$HIRO_{ext}$	The noise cancelling is fantastic. These are the best Headphones I've ever owned. Well worth the money! The batteries last a long time and charge quickly. I have the over the ears pair and also these in the ears.
$HIRO_{sent}$	The Bose QuietComfort 20 headphones offer exceptional noise cancellation and comfort. These Bose QuietComfort 20 headphones are exceptional, providing excellent noise cancellation and sound quality. These headphones are worth the investment due to their exceptional noise-cancelling capabilities. Battery lasts long for noise cancellation, charging quickly. Some users report over 10 hours usage. Some prefer Bose QuietComfort 20 earbuds for flights and noisy environments, others find them dizzying.
$HIRO_{doc}$	The Bose QuietComfort 20 Acoustic Noise Cancelling Headphones are widely praised for their excellent noise cancellation capability, making them a popular choice for frequent flyers and those working in noisy environments. They are also praised for their comfort, sound quality, and long battery life. However, some customers have expressed concerns about the non-replaceable battery and the price. Overall, these headphones are considered a worthwhile investment for their impressive noise cancellation and sound quality.

Table 11: Examples of generated summaries for the 'Bose QuietComfort 20' headphones, from AmaSum. Mistral 7B refers to the opinions of a single user, which is not appropriate for a summary of thousands of reviews.

$q_{1:d}$	(6,2,10)
Evidence	The pool area was very nice. The staff was very Friendly and helpful. the pool area was very enjoyable The pool area is very nice. The pool area is fantastic The pool area was nice the pool area is very attractive The staff was so friendly and helpful. The pool was nice The staff was friendly and accommodating. The pool was very nice and the lobby inviting the staff was welcoming and helpful The hotel staff was very friendly and accommodating. Staff was friendly and helpful. The pool area is very nice and large with several water features. The staff was really great and helpful
$HIRO_{sent}$	The pool area and friendly staff make this hotel a enjoyable stay.
$q_{1:d}$	(10, 8, 6)
Evidence	Free parking is also available to the rear of the hotel. Free parking in the hotel yard. A big plus was the free parking and large selection for continental breakfast. The free plentiful parking (some even under cover) is also one of the better features of the hotel. It has free parking (a big deal in Key West), an awesome pool, and the best free continental breakfast I've ever seen. Parking is free with some covered parking.
$HIRO_{sent}$	Free parking and continental breakfast are notable features of this hotel.
$q_{1:d}$	(9,)
Evidence	I was extremely disappointed in the rooms Only one mirror in the room. The rooms are loud. the rooms are awful the room was small and shabby we were a little disappointed because the room was a lot smaller than we expected Extremely disappointed in the room, although the help was very nice as was the outdoor area. i suspect we were in the older part of the hotel with a double room. Since the parking is directly under the rooms, it was Very loud espcially from 12–3 am.
$HIRO_{sent}$	Rooms were small, loud, and in need of renovation with poor housekeeping.
$HIRO_{doc}$	The Fairfield Inn and Suites Key West received positive reviews for its friendly and helpful staff, attractive and nice pool area, and free parking. However, some guests were disappointed with the small and shabby rooms, lack of storage space, and noise from parking and adjacent rooms. The continental breakfast was also mentioned as a plus.

Table 12: Examples of selected subpaths $q_{1:d}$, the corresponding evidential clusters, the resulting $\text{HIRO}_{\text{sent}}$ output sentences, and the overall HIRO_{doc} summary for a single entity from SPACE. We show only three out of five input clusters, and a subset of all evidence sentences, due to space constraints.