Comparative Opinion Summarization via Collaborative Decoding

Hayate Iso[♌] **Xiaolan Wang**[♌] **Stefanos Angelidis**[♃] **Yoshihiko Suhara**[♌]
[♌]Megagon Labs [♃]University of Edinburgh
{hayate,xiaolan,yoshi}@megagon.ai s.angelidis@ed.ac.uk

Abstract

Opinion summarization focuses on generating summaries that reflect popular subjective information expressed in multiple online reviews. While generated summaries offer general and concise information about a particular hotel or product, the information may be insufficient to help the user compare multiple different choices. Thus, the user may still struggle with the question "Which one should I pick?" In this paper, we propose the *comparative opinion summarization* task, which aims at generating two contrastive summaries and one common summary from two different candidate sets of reviews. We develop a comparative summarization framework CoCoSum, which consists of two base summarization models that jointly generate contrastive and common summaries. Experimental results on a newly created benchmark CoCoTrip show that CoCoSum can produce higher-quality contrastive and common summaries than state-of-the-art opinion summarization models. The dataset and code are available at https://github.com/megagonlabs/cocosum.

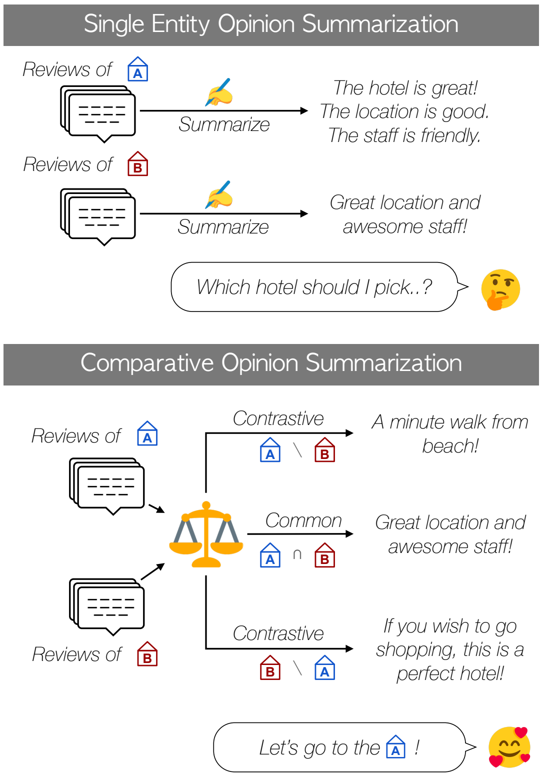

Figure 1: Overview of the comparative opinion summarization task. The model takes two set of reviews about different entities to generate two contrastive opinion summaries, which contain distinctive opinions, and one common opinion summary, which describes common opinions between the two entities.

1 Introduction

Widely available online customer reviews help users with decision-making in a variety of domains (e.g., hotel, restaurant, or company). After creating a list of candidate choices based on initial conditions (e.g., area, price range, restaurant type), the user often has to compare a few choices in depth by carefully reading the reviews to make a final decision. However, it is time-consuming and difficult for the user to detect differences and similarities between the candidates, as those pieces of information are often scattered in different reviews.

The recent success of neural summarization techniques and the growth of online review platforms led to establishing the field of multi-document opinion summarization (Chu and Liu, 2019; Bražinskas et al., 2020b; Amplayo and Lapata, 2020; Suhara et al., 2020; Iso et al., 2021). The goal of multi-document opinion summarization is to generate a summary that represents salient opinions in input reviews of a particular hotel or product, which we refer to as an entity. However, existing opinion summarization techniques are designed to generate a *single-entity opinion* summary that reflects popular opinions for each entity, without taking into account *contrastive and common opinions* that are uniquely (commonly) mentioned in each entity (both entities) as depicted in Figure 1. Therefore, the user still needs to figure out which opinions are

Findings of the Association for Computational Linguistics: ACL 2022, pages 3307 - 3324
May 22-27, 2022

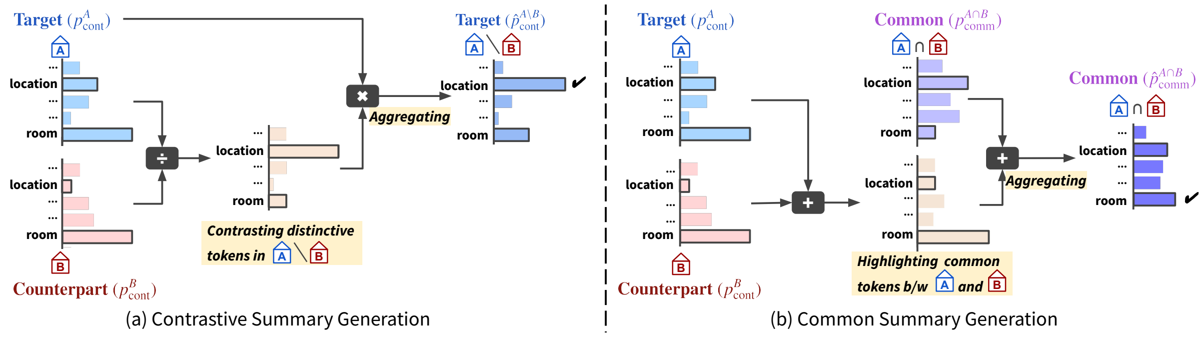

Figure 2: Illustration of Co-decoding: (a) For contrastive summary generation, distinctive words are emphasized by *contrasting* the token probability distribution of target entity against that of the counterpart entity. (b) For common summary generation, entity-pair-specific words are highlighted by *aggregating* token probability distributions of all base models to alleviate the overly generic summary generation issue.

contrastive or common between the entities by carefully reading and comparing summaries generated by existing opinion summarization solutions.

To this end, we take one step beyond the current scope of opinion summarization and propose a novel task of generating contrastive and common summaries by comparing multiple entities, which we refer to as *comparative opinion summarization*. In contrast to the conventional single-entity opinion summarization task that makes a general summary for each entity, the goal of comparative opinion summarization is to generate two contrastive summaries and one common summary from two sets of reviews about two entities. Thus, the user can easily understand distinctive and common opinions about multiple entities. In this paper, we consider pairwise comparison as it is the most common choice and the minimal unit for multiple comparisons.

A key challenge of building a summarizer for the task is that the model has to correctly distinguish what contrastive and common opinions from input reviews of two entities are. Existing opinion summarization models do not implement this functionality as they are designed to summarize popular opinions for a single entity.

To address this issue, we develop a comparative opinion summarization framework CoCoSum, which consists of two base summarization models for contrastive and common opinion summary generation. CoCoSum employs a novel Collaborative Decoding (Co-decoding) algorithm that takes two review sets as input to *compare and contrast* the token probability distributions of the models to generate more distinctive summaries as illustrated in Figure 2.

Experimental results on a newly created comparative opinion summarization benchmark CoCoTrip show that CoCoSum with Co-decoding generate substantially high-quality contrastive and common summaries compared to baseline models including state-of-the-art opinion summarization models.

Our contributions are as follows:

- We propose the novel task of comparative opinion summarization, which takes two review sets as input and outputs two contrastive summaries and one common summary.
- We develop CoCoSum, which consists of two base summarization models and implements a novel Co-decoding algorithm that facilitates generating distinctive and entity-pair-specific summaries by aggregating the token probability distributions of the models.
- We create and release a comparative opinion summarization benchmark CoCoTrip that contains manually written reference summaries for 48 entity pairs.

2 Comparative Opinion Summarization

2.1 Problem Formulation

Let $\mathcal{C}$ be a corpus of reviews on entities from a single domain (e.g., hotels). For each entity e, we define its review set $\mathcal{R}_e = \{r_{e,1}, r_{e,2}, \ldots, r_{e,|\mathcal{R}_e|}\}$.

We define *contrastive opinions* of a target entity A against a counterpart entity B as subjective information that is described only in $\mathcal{R}_A$ but not in $\mathcal{R}_B$ and refer to the summary that contains such contrastive opinions as a *contrastive summary* $y^{A\setminus B}_{\text{cont}}$. Similarly, we define *common opinions* of entities A and B as subjective information that is commonly described in $\mathcal{R}_A$ and $\mathcal{R}_B$ and refer to the summary that contains common opinions as a *common summary* $y^{A\cap B}_{\text{comm}}$. Note that $y^{A\cap B}_{\text{comm}}$ and $y^{B\cap A}_{\text{comm}}$

are identical, thus we consider a single common summary for an entity pair.

We formalize *comparative opinion summarization* as a task to generate two sets of contrastive summaries $y_{\text{cont}}^{A \setminus B}$, $y_{\text{cont}}^{B \setminus A}$, and one common summary $y_{\text{comm}}^{A \cap B}$ from two sets of reviews $\mathcal{R}_A$ and $\mathcal{R}_B$ for a pair of entities A and B. Compared to existing summarization tasks, comparative opinion summarization is the first work that aims to generate abstractive summaries for contrastive and common opinions.

2.2 The CoCoTrip Corpus

As the task requires three types of reference summaries for each *entity pair*, none of the existing benchmarks for single-entity opinion summarization can be used for evaluation. Therefore, we create a comparative opinion summarization corpus CoCoTrip that contains human-written contrastive and common summaries for 48 pairs of entities. We sampled the entity pairs and reviews from the TripAdvisor corpus (Wang et al., 2010).

We sampled 16 reviews for every pair (i.e., 8 reviews for each entity). For every entity pair, we collected 3 gold-standard summaries written by different annotators for two contrastive summaries and one common summary. Details of the corpus creation process are described in Appendix.

We summarize the CoCoTrip dataset and compare it with existing opinion summarization datasets in Table 1. We calculate novel n-grams in gold summaries to evaluate how abstractive/extractive CoCoTrip is. Considering the input and summary length, we confirm that CoCoTrip is sufficiently abstractive compared to the existing opinion summarization datasets.

3 CoCoSum

In order to summarize contrastive and common opinions from two sets of reviews, the comparative opinion summarization task requires the model to compare and contrast two sets of reviews; however, existing single-entity opinion summarization models do not have such functionality. Therefore, we design a "collaborative" decoding solution Co-decoding, which characterizes the target summary distribution by leveraging two base summarization models.

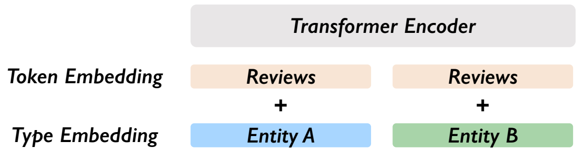

Figure 3: Encoder of the base common summarization model has *type embeddings* to distinguish the original entity.

3.1 Base Summarization Model

CoCoSum consists of two base summarization models. The base contrastive summarization model is a single-entity summarization model that takes only reviews of the target entity as input, while the base common summarization model takes reviews of two entities as input. In both cases, the input reviews are concatenated into a single sequence before encoding. To help the encoder distinguish the entity, we introduce additional *type embeddings* into the input layer of the encoder as shown in Figure 3.

For common summary generation (i.e., $y_{\text{comm}}^{A \cap B} = y_{\text{comm}}^{B \cap A}$), the model should generate the same common summary for the same entity pair regardless of the input order of review sets. Thus, we create two input sequences (i.e., $A \cap B$ and $B \cap A$) and merge the token probability distributions of the two sequences for a summary generation.

3.2 Collaborative Decoding

As illustrated in Figure 2, Co-decoding combines predictions of the target and the counterpart (and common, for common summary generation) opinion summarization models during the inference time. The key idea of Co-decoding is to aggregate token probability distributions of contrastive summarization model $p_{\text{cont}}(\cdot)$ and common summarization model $p_{\text{comm}}(\cdot)$ at each step, so the two models can collaboratively generate (1) contrastive summaries that contain distinctive opinions that do not appear in the counterpart review set and (2) common summaries that only contain common opinions that appear in both target and counterpart review sets.

Contrastive summary generation To improve the distinctiveness of generated contrastive summaries that only contains entity-specific opinions, we consider *penalizing* the tokens that are likely to appear in the counterpart entity. That is, we use two token probability distributions and highlight tokens that are distinctive compared to the coun-

	Task	Input length	Summary length	Domain	% of novel n-grams in gold summary unigram	bigram	trigram	4-gram
CoCoTrip (Ours)	Contrastive	1529.4	132.9	Hotels	22.81	72.41	91.43	97.08
	Common		20.3		9.27	51.75	84.52	95.75
Chu and Liu (2019)	Single	581.1	70.4	Businesses	30.87	83.23	96.60	99.18
Bražinskas et al. (2020b)	Single	473.4	59.8	Products	26.23	77.52	93.24	97.43
Angelidis et al. (2021)	Single	16160.6	83.6	Hotels	1.98	21.13	47.14	63.86

Table 1: Statistics of CoCoTrip and other benchmarks. CoCoTrip has a comparable corpus size against the benchmarks while offering unique characteristics (i.e., three types of reference summaries for a pair of entities). The average input length in tokens is calculated using concatenated input reviews.

terpart entity by using the *token ratio distribution* between them. We also introduce a trade-off hyperparameter δ that controls the balance between the original token distribution and the token ratio distribution:

$$\hat{p}^{A\setminus B}_{\text{cont}}(y_t) \propto p^A_{\text{cont}}(y_t)\left(\frac{p^A_{\text{cont}}(y_t)}{p^B_{\text{cont}}(y_t)}\right)^{\delta}, \qquad (1)$$

where we denote the token probability distribution of the base contrastive summarization model given the previously generated tokens $y_{<t}$ and the set of input reviews $\mathcal{R}_e$ for entity $e \in \{A, B\}$ at t-th step by $p^e_{\text{cont}}(y_t) := p_{\text{cont}}(y_t \mid y_{<t}, \mathcal{R}_e)$. Note that for both $p^A_{\text{cont}}(y_t)$ and $p^B_{\text{cont}}(y_t)$, we use the same prefix $y_{<t}$. For the other contrastive summary $\hat{y}^{B\setminus A}_{\text{cont}}$, the token probability can be obtained by swapping A and B in Eq. (1).

Co-decoding for contrastive summary generation is illustrated in Figure 2 (a). The intuition behind this approach is that the token ratio distribution $\frac{p^A_{\text{cont}}(y_t)}{p^B_{\text{cont}}(y_t)}$ (i.e., $A \wedge \neg B$) highlights distinctive tokens that are relatively unique to the target entity, which are emphasized by combining with the original token distribution. This can be considered a variant of Product-of-Experts (PoE) (Hinton, 2002; Liu et al., 2021), which models Logical AND with multiple probabilistic distributions.

Common summary generation Common summaries should contain common opinions that are about a given pair of entities. However, we observe that simply fine-tuned summarization models tend to generate overly generic summaries that can be true for any entity pair.

To incorporate the entity-specific information into the common summary, we design Co-decoding to use the sum of the token probability distributions of the contrastive summarization model, which is then combined with the original token probability distribution using a trade-off hyperparameter γ:

$$\hat{p}^{A\cap B}_{\text{comm}}(y_t) \propto p^{A\cap B}_{\text{comm}}(y_t) + \gamma \sum_{E\in\{A,B\}} p^E_{\text{cont}}(y_t), \qquad (2)$$

where we denote the token probability distribution of the base common summarization model by $p^{A\cap B}_{\text{comm}}(y_t) := p_{\text{comm}}(y_t \mid y_{<t}, \mathcal{R}_A, \mathcal{R}_B)$.

Co-decoding for common summary generation is illustrated in Figure 2 (b). The intuition behind this approach is that we first identify salient tokens for the input entity pair by adding the token probability distributions of contrastive summaries: $p^A_{\text{cont}}(y_t) + p^B_{\text{cont}}(y_t)$ (i.e., $A \vee B$), which is then combined with the original distribution using the trade-off hyperparameter γ. This can be considered a variant of Mixture-of-Experts (MoE) (Jacobs et al., 1991), which models Logical OR with multiple probabilistic distributions and is suitable for *interpolating* the token probability distribution of models with different characteristics.

We would like to emphasize that Co-decoding is a token probability distribution calculation method for comparative opinion summarization based on two summarization models; thus, it is flexible of the choice of the base summarization model and the decoding algorithm.

4 Evaluation

4.1 Experimental Settings

We build two versions of CoCoSum using self-supervised training (Self-supervised) and few-shot learning (Few-shot). We evaluate the summarization performance of the two versions with and without Co-decoding. For all the base models, we use a pre-trained LED model (Beltagy et al., 2020), which uses sparse attention to handle long sequences and thus is suitable for the purpose.[1]

[1] https://huggingface.co/allenai/led-base-16384

	Contrastive R1	R2	RL	BS	Common R1	R2	RL	BS	Pair DS
Self-supervised									
Extaractive models									
LexRank$_{\text{TFIDF}}$ (Erkan and Radev, 2004)	35.38	7.39	18.25	22.61	22.51	4.00	15.26	24.65	63.28
LexRank$_{\text{BERT}}$ (Reimers and Gurevych, 2019)	32.65	5.67	16.67	20.51	17.91	2.95	12.60	24.83	65.56
Abstractive models									
MeanSum (Chu and Liu, 2019)	34.19	7.84	19.76	23.89	13.09	0.85	10.41	16.13	65.98
OpinionDigest (Suhara et al., 2020)	37.30	8.67	20.36	21.77	21.52	4.41	15.26	17.06	64.87
CopyCat (Bražinskas et al., 2020b)	35.30	8.39	18.64	21.91	36.16	11.91	25.15	50.16	40.80
BiMeanVAE (Iso et al., 2021)	37.44	9.41	22.02	24.33	38.47	14.17	27.46	50.98	42.55
CoCoSum (Ours)									
Self-supervised	40.01	**10.80**	**21.97**	**30.02**	**41.13**	**15.25**	**30.60**	**54.65**	**66.00**
w/o Co-decoding ($\delta = \gamma = 0.$)	**40.78**	10.66	21.53	29.90	40.40	14.13	29.81	54.28	57.63
Few-shot	42.22	12.11	24.13	**35.63**	**46.80**	**20.68**	**35.62**	**61.52**	**74.02**
w/o Co-decoding ($\delta = \gamma = 0.$)	**43.65**	**12.83**	**24.93**	35.42	45.90	19.59	34.40	59.32	71.69
Human upper bound	47.37	13.00	26.03	37.69	52.26	19.16	39.89	61.10	71.79

Table 2: ROUGE and BERT scores (summarization quality) for contrastive and common summaries on CoCoTrip and the distinctiveness score (DS) of generated summaries. CoCoSum significantly improves the distinctiveness while keeping high summarization quality. Human upper bound is measured by calculating the corresponding score across multiple reference summaries.

For self-supervised training, we use the TripAdvisor review corpus (Wang et al., 2010) to construct pseudo review-summary pairs following Elsahar et al. (2021) with two modifications: 1) We filter reviews with different word length ranges for contrastive ($[100, 150]$) and common ($[15, 50]$) base models to accommodate the different average summary lengths. 2) For the self-supervised base common summarization model, as it takes two sets of reviews (i.e., $\mathcal{R}_A$, $\mathcal{R}_B$) as input, we retrieve and merge review-summary pairs, based on the summary similarity, to make a pseudo training dataset.

For few-shot learning, we use 20 instances of CoCoTrip for further fine-tuning self-supervised base summarization models. Detailed analysis of the few-shot learning strategies can be found in Appendix.

For evaluation, we used the remaining 10 instances of CoCoTrip for development and 18 instances for testing.

For Co-decoding, we used top-p vocabulary (Holtzman et al., 2020), which is the smallest token set whose cumulative probability exceeds p, with $p = 0.9$ for $p^{A}_{\text{cont}}(y_t)$, $p^{B}_{\text{cont}}(y_t)$, and $p^{A\cap B}_{\text{comm}}(y_t)$. We used Beam Search with a width of 4. We chose δ and γ using the dev set.

To access the quality of CoCoSum, we evaluated the performance of a variety of baseline approaches:

LexRank$_{\text{TFIDF}}$ (Erkan and Radev, 2004): The classic unsupervised opinion summarization solution;
LexRank$_{\text{BERT}}$ (Erkan and Radev, 2004; Reimers and Gurevych, 2019): LexRank approach with Sentence BERT (Reimers and Gurevych, 2019) embeddings[2];
MeanSum (Chu and Liu, 2019): the unsupervised single entity opinion summarization solution[3];
CopyCat (Bražinskas et al., 2020b): a single entity opinion summarization solution based on leave-one-out reconstruction[4];
BiMeanVAE (Iso et al., 2021): an optimized single entity opinion summarization solution[5] for MeanSum.

For those baseline models above, we use $\mathcal{R}_A$ (or $\mathcal{R}_B$) as input for the contrastive summary and both $\mathcal{R}_A$ and $\mathcal{R}_B$ as input for the common summary.

OpinionDigest (Suhara et al., 2020): a weakly supervised opinion summarization approach.[6] We customize OpinionDigest for comparative opinion summarization. Specifically, we categorize opinion clusters extracted from $\mathcal{R}_A$ and $\mathcal{R}_B$ as "contrastive" if the cluster only contains opinions from a single entity and "common" if the cluster contains opinions from both of the entities. In this way, OpinionDigest can leverage the extracted

[2] https://github.com/UKPLab/sentence-transformers
[3] https://github.com/sosuperic/MeanSum
[4] https://github.com/abrazinskas/Copycat-abstractive-opinion-summarizer
[5] https://github.com/megagonlabs/coop
[6] https://github.com/megagonlabs/opiniondigest

	Content overlap			Content support			Quality		
	Overlap ↓	Partial ↓	Not ↑	Full ↑	Partial ↑	No ↓	Coh ↑	Info ↑	Non-red ↑
BiMeanVAE	64.45	20.19	15.35	45.23	31.54	23.24	3.78	2.34	3.11
OpinionDigest	20.73	21.15	58.12	42.31	28.53	29.17	3.53	2.28	3.29
CoCoSum$_{\text{few}}$	**4.80**	**25.20**	**70.00**	**63.50**	**24.09**	**12.41**	4.10	**2.81**	**4.38**
w/o Co-decoding	10.02	22.14	67.84	58.27	25.90	15.83	**4.19**	2.80	4.34

Table 3: Human evaluations on content overlap, content support, coherence (coh.), informativeness (info.), and non-redundancy (non-red).

opinion clusters to generate contrastive and common summaries.

4.2 Automatic Evaluation

Evaluation metrics For summarization quality, we use ROUGE 1/2/L F1 scores (Lin, 2004)[7] and BERTScore (Zhang et al., 2020)[8] as automatic evaluation based on reference summaries.

To evaluate the *distinctiveness* of generated summaries, we calculate the average distinctiveness score (DS) between generated contrastive summaries and common summaries for all entity pairs defined as follows:

$$\mathrm{DS} = 1 - \frac{\sum_{(y,z)\in\hat{\mathcal{Y}}^{(2)}} |\mathcal{W}_y \cap \mathcal{W}_z| - 2|\bigcap_{y\in\hat{\mathcal{Y}}} \mathcal{W}_y|}{|\bigcup_{y\in\hat{\mathcal{Y}}} \mathcal{W}_y|},$$

where $\hat{\mathcal{Y}} := \{\hat{y}_{\text{cont}}^{A\setminus B}, \hat{y}_{\text{cont}}^{B\setminus A}, \hat{y}_{\text{comm}}^{A\cap B}\}$, $\mathcal{W}_y$ is the token bag of generated summary $y \in \hat{\mathcal{Y}}$, and $\hat{\mathcal{Y}}^{(2)}$ is the 2-subsets of $\hat{\mathcal{Y}}$. The DS will be higher if the word overlaps between two generated contrastive summaries $\hat{y}_{\text{cont}}^{A\setminus B}$, $\hat{y}_{\text{cont}}^{B\setminus A}$, and a generated common summary $\hat{y}_{\text{comm}}^{A\cap B}$ are smaller.

Results As shown in Table 2, CoCoSum outperforms the baseline methods for the ROUGE and BERT scores (summarization quality) and the distinctiveness score (DS), showing the effectiveness of our self-supervised dataset and Co-decoding. Comparing the summarization quality by CoCoSum and CoCoSum w/o Co-decoding, we confirm that Co-decoding significantly improves the distinctiveness especially in self-supervised setting while maintaining the summarization performance.

Among the baseline methods, BiMeanVAE shows the highest ROUGE scores while performing poorly for the distinctiveness score. Although MeanSum and OpinionDigest show high distinctiveness scores, they show significantly worse performance on the common summary generation task. The results indicate it is challenging for existing opinion summarization models to improve the distinctiveness of generated summaries while keeping them high-quality for both of the tasks.

4.3 Human Evaluation

For human evaluation, we hired contractors from Upwork[9] platform and conducted three sets of human evaluation comparing CoCoSum with two representative baselines—BiMeanVAE and OpinionDigest.

First, we asked the human annotators to evaluate the overlapped content between the contrastive summaries and the common summary for a given entity pair. More specifically, for every sentence in the summary, we asked human annotators to judge if its content is *overlap*, *partially overlap*, or *not overlap* with the other two summaries. According to the problem formulation, less overlap, i.e., not or partially overlap, is preferred. As shown in Table 3, CoCoSum is significantly better than CoCoSum w/o Co-decoding, and is substantially better than BiMeanVAE and OpinionDigest. This result also aligns with our automatic evaluation on the distinctiveness (DS in Table 2), and it demonstrates that CoCoSum can produce more *distinctive* contrastive and common summaries.

Second, we conducted a summary content support study to evaluate how faithful the generated summaries are toward the input reviews. Similar to content overlap, for every sentence in the summary, we asked human annotators to judge if its content is *fully supported*, *partially supported*, or *not supported* by the corresponding input review sentences. Note that the input review sentences are selected based on sentence-level labels we acquired from CoCoTrip. The results show that CoCoSum is able to generate the most faithful

[7] https://github.com/Diego999/py-rouge

[8] DeBERTa NLI model (He et al., 2021) and baseline rescaling are used.

[9] https://www.upwork.com

summaries compared to all the other baselines.

Lastly, we asked the human annotators to give ratings (from 1 to 5) for the generated summaries with respect to three criteria, namely *coherence*, *informativeness*, and *non-redundancy*. We report the average ratings (Harpe, 2015) for the summaries generated from different methods in Table 3. As shown in the table, summaries generated by CoCoSum is slightly less coherent than CoCoSum w/o Co-decoding. This slight degradation is expected because Co-decodingadjusts the token probability to encourage contrastive/common content, thus it may also prioritize tokens that are less coherent. Other than coherence, CoCoSum shows slightly better informativeness and non-redundancy. Meanwhile, compared to BiMeanVAE and OpinionDigest, CoCoSum shows much better performance on all the three criteria.

5 Analysis

5.1 Distinctiveness in Generated Summaries

In addition to the summarization quality, distinctiveness is another important factor for comparative opinion summarization to help the user pick one against the other. Therefore, we conduct additional analysis to investigate the quality of distinctiveness in generated summaries.

How distinctive are generated contrastive summaries for each entity pair? To complement our experiments on the distinctiveness score (in Table 2), which considers both types of generated summaries, we further evaluate *intra-entity-pair BERTScore (Intra-BERTScore)* only between two contrastive summaries for each entity pair to measure the *intra-entity-pair distinctiveness* defined by the average of BERTScore($\hat{y}_{\text{cont}}^{A\setminus B}$, $\hat{y}_{\text{cont}}^{B\setminus A}$).

Figure 4 shows that in both self-supervised and few-shot settings, CoCoSum significantly outperforms the state-of-the-art opinion summarization model (BiMeanVAE). The results confirm that Co-decoding successfully generates more distinctive opinions of each other, and the hyperparameter δ controls the trade-off between the summarization quality (BERTScore) and the distinctiveness (Intra-BERTScore).

5.2 Analysis on Co-decoding Design

Our design of Co-decoding uses different types of distribution aggregation methods for contrastive (Eq. (1)) and common summary generation (Eq. (2)). To support those intuitive designs, we examine how the quality of generated summaries is affected when different configurations in Co-decoding are used for each task. The full table is presented in the Appendix.

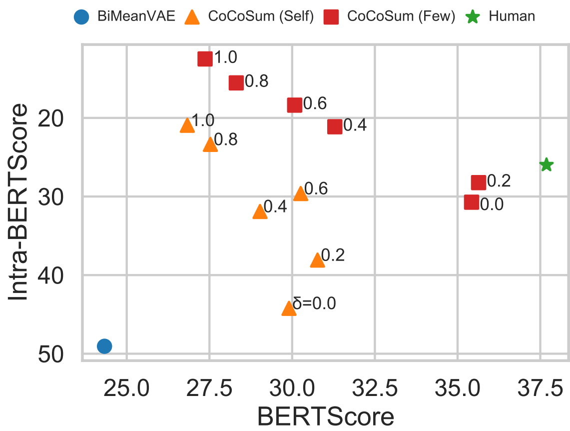

Figure 4: BERTScore and Intra-BERTScore for generated contrastive summaries with different hyperparameters δ. The goal is to generate high quality and distinctive summaries (upper right).

Contrastive summary generation First, we tested the MoE style aggregation that is used for contrastive summary generation. Specifically, we use addition to combine the original distribution and the ratio distribution instead of multiplication: $p_{\text{cont}}^{A\setminus B}(y_t) + \delta\left(p_{\text{cont}}^{A\setminus B}(y_t)/p_{\text{cont}}^{B\setminus A}(y_t)\right)$.

With this configuration, we observe significant degradation of summarization quality (e.g., 3.14 on R1) due to a serious distribution collapse issue in the aggregated token probability distribution. This is mainly caused by the lack of the *cancellation effect* obtained by the PoE style aggregation. That is, if the probability of a token were low in the ratio distribution, it would be canceled out via the *multiplication* operation.

We also tested another way to highlight contrastive opinions using the common summary generation model for the ratio distribution. That is, we replace the ratio distribution in Eq. (1) with $p_{\text{cont}}^{A}(y_t)/p_{\text{comm}}^{A\cap B}(y_t)$.

This configuration shows competitive performance as the original CoCoSum in both self-supervised and few-shot settings, supporting the effectiveness of Co-decoding regardless of the specific model choice. However, this configuration requires an additional base common opinion summarization model $p_{\text{comm}}^{A\cap B}$. Thus, we decided to use

CoCoSum	Intra-ROUGE1/2/L = (**36.84, 8.51, 23.16**)
Entity ID: 305813	**Entity ID: 305947**
The Langham Place Hotel is a 4-star hotel. It is centrally located **and has easy access to the mall and cinema next door. The room was lovely with a great view. The bed in the room was firm and not too comfy. The spa facilities here at the hotel are of a really high standard.** The staff at this hotel are generally excellent and very much co-operative. **The hotel had over priced buffet meals and snacks and drinks but the club floor facilities are of such a high standard that you know you are worth it.**	**The Metropark Kowloon is a good hotel to stay at for a week or longer. It's ideally located for those who needs to shower and hit the bed after a full day of sightseeing/shopping.** The rooms in the hotel are clean, modern **and air-conditioning works well. The food served in the restaurant was varied and varied. The hotel provided a free shuttle service to Mongkok and the harbour area. The Ladies Market in Mong Kok is a pleasant walk away but the hotel bus route takes us close by.**

CoCoSum w/o Co-decoding	Intra-ROUGE1/2/L = (45.03, 11.64, 26.18)
Entity ID: 305813	**Entity ID: 305947**
The hotel has a great spa and sauna facilities and is centrally located to other attractions. **It is also worth booking into the club floor for the daily cocktail hour** and internet access. *The hotel could not do enough for you. The staff at the hotel were not very much co-operative and could not help enough.* **The spa facilities were of a very high standard** and the food was of a really good quality. *The taxi drivers take advantage of the hotel's direct access to the mall and cinema next door.*	**The Metropark Kowloon is a good hotel to stay at for a week or longer.** The hotel is in a good location *and only a short walk away from the main shopping areas of Hong Kong.* The rooms in the hotel are clean **and with air conditioning but the rooms can be quite chilly compared to the humdity outside.** The staff at the hotel were very helpful and accommodating. **The buffet breakfast was really good and varied. The food served in the restaurant was really varied and tasty.** The Sip Sip bar offered a great variety of cocktails.

Table 4: Contrastive summaries generated by CoCoSum with and w/o Co-decoding for an example entity pair. Distinctive (common) opinions are highlighted in **blue** (magenta), and hallucinated content is in italics.

the simpler configuration Eq. (1) as the default setting.

Common summary generation Similarly, we verified the effectiveness of the PoE style configuration for common summary generation. That is, we use multiplication instead of addition: $p^{A\cap B}_{\text{comm}}(y_t) \prod_{E\in\{A,B\}} p^{E}_{\text{cont}}(y_t)^{\gamma}$.

This configuration consistently under-performs with the original Co-decoding for both summarization and inter-distinctiveness scores. This indicates that PoE focuses too much on the tokens that are likely to appear in both contrastive and common summaries, and thus it tends to generate overly generic summaries.

5.3 Qualitative Analysis

Does Co-decoding generate more distinctive opinions? Table 4 shows example generations by CoCoSum with and w/o Co-decoding for contrastive summary generation. While both models generate summaries that are consistent with the target entity reviews, the summaries generated by CoCoSum w/o Co-decoding tend to contain common opinions that are true for both of the entities and are against the purpose of comparative opinion summarization. On the contrary, CoCoSum actively generates opinions that can only be generated by the target entity's model p^{A}_{cont}, and thus the generated summary contains more **contrastive opinions** for users to compare the entities.

Do different pairs yield different summaries? Distinctive opinions can change when the entity to be compared changes. Table 5 shows the generated contrastive summaries using different entities as counterpart. As in the previous example, CoCoSum can generate generally consistent summaries with the target entity reviews in each setting, but also it uses **different opinions** to generate summaries. In other words, the model can highlight different opinions by comparing them with different entities, and thus generate summaries that include significantly different opinions for each.

6 Related Work

Abstractive opinion summarization aims to generate a fluent summary that reflects salient opinions in input reviews. Due to the lack of sufficient amount of reference summaries, the most common solution is the unsupervised approach (Chu and Liu, 2019; Bražinskas et al., 2020b; Amplayo and Lapata, 2020; Suhara et al., 2020; Amplayo et al., 2021; Iso et al., 2021; Elsahar et al., 2021; Im et al., 2021; Wang and Wan, 2021; Isonuma et al., 2021; Ke et al., 2022, *inter alia*).

Target Entity ID: 614392 vs Counterpart Entity ID: 1022738
The Pullman hotel is ideally situated for a city/beach vacation. **Port Olympic,** the Beach **and Barcelonetta** are all within walking distance. **The hotel has 2 great pools, one roof top pool with bar** and *one rooftop pool with a bar.* **It's not cheap though. The fitness centre in the hotel is tiny but there is a fitness park about 2 minutes walking distance for 16eur/day which provide a good facility.**
Target Entity ID: 614392 vs Counterpart Entity ID: 256595
The Pullman Hotel Barcelona is a stylish hotel next to the beach **with impeccable customer service.** The hotel is well situated in Barcelona, **not too far from the 5 star establishment, the Arts Hotel etc. The rooms in the hotel are of a good size and nicely decorated. The room has a great balcony and sea view and the bed is incredibly comfortable. The bathroom is also really luxurious. The staff at the hotel were really attentive and really go out of their way to treat all of their guests like they are royalty.The Mini bar was expensive so avoid at all costs FYI.**

Table 5: Contrastive summaries (Entity ID: 614392) generated by COCOSUM with Co-decoding using different entities as counterpart (Entity ID: 1022738 and 256595). The COCOSUM can generate completely different summaries by different conditioning. **Different opinions summarized** are color-coded and *hallucinated content* is in *italics*

Recent opinion summarization models use the few-shot learning approach that fine-tunes a pre-trained Transformer model with a limited amount of pairs of input reviews and reference summaries. Bražinskas et al. (2020a) and Oved and Levy (2021) show that the few-shot learning approach substantially outperforms unsupervised learning models.

All the existing methods listed above are designed for general opinion summarization and, thus, are not necessarily suitable for comparative opinion summarization, as shown in the experiments.

Comparative summarization There is a line of work on extracting comparative information from single/multiple documents. Lerman and McDonald (2009) defined the contrastive summarization problem and presented early work on the problem. Their method selects sentences so that two sets of summaries can highlight differences. Wang et al. (2013) developed an extractive summarization method for a problem of Comparative Document Summarization, which is to select the most discriminative sentences from a given set of documents. Bista et al. (2019) tackled a similar problem by selecting documents that represent in-cluster documents while they are useful to distinguish from other clusters.

Other studies (Kim and Zhai, 2009; Huang et al., 2011; Sipos and Joachims, 2013; Ren et al., 2017) tackled similar tasks by developing extracting sentences/phrases from given sets of documents for comparative document analysis. Topic models have also been used to capture comparative topics for better understanding text corpora, but they do not generate textual summaries (Ren and de Rijke, 2015; He et al., 2016; Ibeke et al., 2017).

Our work differs from the existing work in two points. First, none of them focuses on generating common summaries. Second, all of the previous studies for contrastive summary generation use the extractive approach. To the best of our knowledge, we are the first to develop an opinion summarization model and a benchmark for the abstractive contrastive and common summary generation tasks.

7 Conclusions

In this paper, we propose a new comparative opinion summarization task, which aims to generate contrastive and common summaries from reviews of a pair of entities, to help the user answer the question "Which one should I pick?" To this end, we develop a comparative summarization framework COCOSUM, which consists of two base summarization models; COCOSUM also implements Co-decoding, which jointly uses the token probability distribution of each model to generate more distinctive summaries in the decoding step.

For evaluation, we created a comparative opinion summarization benchmark COCOTRIP based on the TripAdvisor review corpus. Experimental results on COCOTRIP show that COCOSUM with Co-decoding significantly outperforms existing opinion summarization models with respect to both summarization quality and distinctiveness. We also confirm that Co-decoding successfully augments COCOSUM, so it can generate more distinctive contrastive and common summaries than other models through comprehensive analysis.

Acknowledgements

We thank Wang-Chiew Tan for valuable inputs that helped create initial ideas and the anonymous reviewers for their insightful comments.

References

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

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

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.

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

Umanga Bista, Alexander Mathews, Minjeong Shin, Aditya Krishna Menon, and Lexing Xie. 2019. Comparative document summarisation via classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 20–28.

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

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

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

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China. Association for Computational Linguistics.

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

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

Spencer E Harpe. 2015. How to analyze likert and other rating scale data. *Currents in pharmacy teaching and learning*, 7(6):836–850.

Lei He, Wei Li, and Hai Zhuge. 2016. Exploring differential topic models for comparative summarization of scientific papers. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1028–1038, Osaka, Japan. The COLING 2016 Organizing Committee.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Geoffrey E. Hinton. 2002. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

Xiaojiang Huang, Xiaojun Wan, and Jianguo Xiao. 2011. Comparative news summarization using linear programming. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 648–653, Portland, Oregon, USA. Association for Computational Linguistics.

Ebuka Ibeke, Chenghua Lin, Adam Wyner, and Mohamad Hardyman Barawi. 2017. Extracting and understanding contrastive opinion through topic relevant sentences. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 395–400, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403, Online. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex Aggregation for Opinion Summarization. In *Findings of the Association for Computational Linguistics:*

EMNLP 2021, pages 3885–3903, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2021. Unsupervised abstractive opinion summarization by generating sentences with tree-structured topic guidance. *Transactions of the Association for Computational Linguistics*, 9(0):945–961.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.

Wenjun Ke, Jinhua Gao, Huawei Shen, and Xueqi Cheng. 2022. Consistsum: Unsupervised opinion summarization with the consistency of aspect, sentiment and semantic. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 467–475, New York, NY, USA. Association for Computing Machinery.

Hyun Duk Kim and ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 385–394, New York, NY, USA. Association for Computing Machinery.

Kevin Lerman and Ryan McDonald. 2009. Contrastive summarization: An experiment with consumer reviews. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 113–116, Boulder, Colorado. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR*.

Nadav Oved and Ran Levy. 2021. PASS: Perturb-and-select summarizer for product reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 351–365, Online. Association for Computational Linguistics.

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.

Xiang Ren, Yuanhua Lv, Kuansan Wang, and Jiawei Han. 2017. Comparative document analysis for large text corpora. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 325–334, New York, NY, USA. Association for Computing Machinery.

Zhaochun Ren and Maarten de Rijke. 2015. Summarizing contrastive themes via hierarchical non-parametric processes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 93–102, New York, NY, USA. Association for Computing Machinery.

Ruben Sipos and Thorsten Joachims. 2013. Generating comparative summaries from reviews. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 1853–1856, New York, NY, USA. Association for Computing Machinery.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2013. Comparative document summarization via discriminative sentence selection. *ACM Trans. Knowl. Discov. Data*, 7(1).

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 783–792, New York, NY, USA. Association for Computing Machinery.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

	Abst.	Cont.	Comm.
Chu and Liu (2019)	✓		
Bražinskas et al. (2020a,b)	✓		
Lerman and McDonald (2009)		✓	
Huang et al. (2011)		✓	
Sipos and Joachims (2013)		✓	
Ren et al. (2017)†		✓	✓
This work	✓	✓	✓

Table 6: Novelty of comparative opinion summarization against existing (opinion) summarization tasks. This work is the first task that targets to generate abstractive summaries (Abst.) for contrastive (Cont.) and common (Comm.) opinions. Note that Ren et al. (2017) extract keywords instead of creating textual summary.

A Comparative Opinion Summarization

Table 6 shows the task comparison against existing summarization tasks. Comparative opinion summarization is the first work that aims to generate abstractive summaries for contrastive and common opinions.

B The CoCoTrip Corpus

B.1 Entity-Pair Selection

For comparative opinion summarization, each of the selected entity pairs should always be comparable. To achieve this goal, we leverage the meta information of hotels in the TripAdvisor corpus to make sure that the selected entity pairs always locate in the same region (e.g., Key West of Florida).

B.2 Annotation

The input for each entity pair includes 16 reviews, which may be too difficult for human writers to write summaries from. Thus, we used a two-stage annotation method to ensure the quality of reference summaries.

Sentence Annotation Our first annotation task focuses on obtaining a set of sentences that contain contrastive and common opinions. Since the average number of sentences in each entity pair (90 in CoCoTrip) was too many to annotate at once, we grouped sentences based on their aspect category to further simplify the annotation task, In particular, we first split input reviews into sentences. Then, we grouped sentences into 6 aspect categories (i.e., general, staff, food, location, room, and others) using a BERT-based aspect category classifier trained with 3K labeled sentences. By doing so, we ensure that the number of sentences annotators need to review each time is no more than 20. For every sentence from entity e_A (e_B), we asked human annotators to compare it against a group of reference sentences of the same aspect category from entity e_B (e_A) and to distinguish whether it contains any common opinions that also appear in the reference sentences.

For the sentence annotation task, we hired 6 annotators from Appen's[10] expert worker pool with a cost of \$0.85 per annotation. We collected 3 annotations for each review and finalized the label through a majority vote. We obtained labels suggesting whether it contains contrastive or common opinions for every sentence in the entity pairs with the sentence annotation task. The inter annotator agreement (Fleiss' kappa) is 0.5048. The task interface is shown in Figure 5.

Summary Collection In the second annotation task, we first asked human writers to write aspect-based summaries. To exclude unreliable labels obtained in the previous step, we displayed two sets of sentences, one from each entity, to human writers for the summary collection task. This helps human writers ignore irrelevant or incorrectly labeled sentences. For example, to obtain the contrastive summary for aspect location, we first show two corresponding sets of contrastive sentences from both e_A and e_B based on the labels we collected in the previous annotation step. Then, we asked human writers to write two contrastive summaries for e_A and e_B, respectively. Similarly, we asked human writers to write a single common summary by showing two corresponding sets of common sentences. By doing so, we obtained aspect-based summaries for each entity pair, which are then concatenated into a reference summary.

Similar to the sentence annotation task, we also hired workers from Appen's expert worker pool. We hired 4 expert workers for the task with an hourly rate of \$18. For contrastive (common) summaries, annotators requires in average 208 (107) seconds to complete a summary. For every entity pair, we collected 3 reference summaries for each of *two* contrastive summary generation and *one* common summary generation tasks. The task interface is shown in Figure 6. Since it is a text summarization task, we report their agreement via ROUGE/BERTScore in Table 2 as *Human upper bound*. As shown, annotators acquires 47.37

[10] https://appen.com/

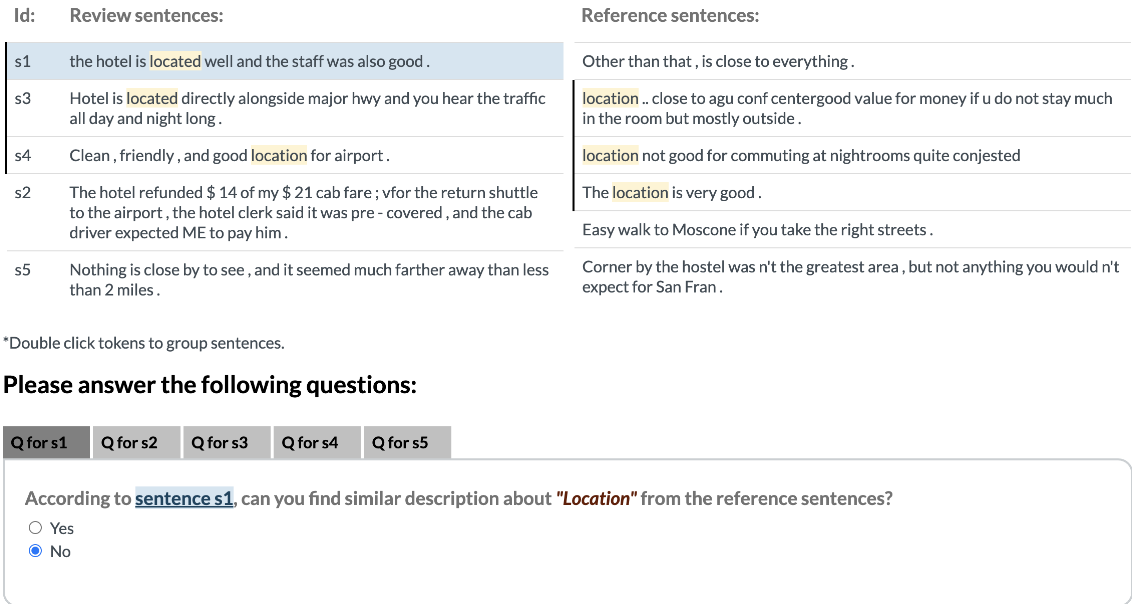

Figure 5: Sentence Annotation Task. By showing sentences of the same aspect category, it is easier for annotators to compare two group of sentences (from two entities). To further facilitate the annotation process, we also provide several additional features, such as allowing workers to group sentences that contain the same token through double clicking, and to highlight sentences through hovering over the sentence label.

ROUGE-1 and 37.69 BERTScore, both are significantly higher than the baseline approaches.

C Additional Experimental Details

C.1 Training details

Major hyper-parameters for training models are reported in Table in 9 and 10 following the "Show-You-Work" style suggested by Dodge et al. (2019).

C.2 Training Dataset for Self-Supervision

We collected synthetic reviews-summary pairs from the TripAdvisor review corpus for self-supervised training. Algorithm 1 shows the review summary pair collection procedure, which is based on Elsahar et al. (2021) with a few modifications.

D Additional Evaluation Results

D.1 Analysis on Few-shot learning design

To explore the best few-shot learning design, we tested three different learning strategies, SELF-THEN-FEW, MULTI-TASKING, and ONLY FEW-SHOT. The SELF-THEN-FEW strategy further fine-tunes the self-supervised summarization model by few-shot training examples. The MULTI-TASKING strategy is to train the summarization model with self-supervised data and the few-shot data jointly. The ONLY FEW-SHOT only fine-times a transformer model initialized with a pre-trained checkpoint, which is the `led-base-16384` in our case.

The experimental results show in Table 8, and we found that while the ONLY-FEW-SHOT configuration shows surprisingly performs well compared to the MULTI-TASKING, the SELF-THEN-FEW strategy performs generally well both on contrastive and common opinion summarizations. Thus, we adapt the SELF-THEN-FEW to build the base summarization models for our experiments.

Algorithm 1 The algorithm for building synthetic training dataset $\mathcal{C}_{\text{synthetic}}$

Input: Raw review sets $\mathcal{C}_{\text{raw}} = \{\mathcal{R}_e\}_{e \in \mathcal{E}}$, task $T \in \{\texttt{contrastive}, \texttt{common}\}$, number of input reviews n, number of synthetic data size K

Output: Synthetic reviews-summary pairs $\mathcal{C}_{\text{synthetic}}$ for task T

procedure BUILDTRAIN($\mathcal{C}_{\text{raw}}$, T, n)
    set synthetic dataset $\mathcal{C}_{\text{synthetic}} \leftarrow \{\}$
    for all review set $\mathcal{R}_e \in \mathcal{C}_{\text{raw}}$ **do**
        for all review $r \in \mathcal{R}_e$ **do**
            $\mathcal{R}_{e,r} := \{r_1, \dots, r_n\} = \underset{\substack{\mathcal{R}'_e \subset \mathcal{R}_e \setminus \{r\} : |\mathcal{R}'_e| = n, \\ \forall r_i \in \mathcal{R}'_e : 50 \le |r_i| \le 150}}{\operatorname{argmax}} \sum_{i \in \mathcal{R}'_e} \text{sim}(r, r_i)$
            if $T = \texttt{contrastive}$ & length of r is between 100 and 150 **then**
                $\mathcal{C}_{\text{synthetic}} \leftarrow \mathcal{C}_{\text{synthetic}} \cup \{(\mathcal{R}_{e,r}, r)\}$
            else if $T = \texttt{common}$ & length of r is between 15 and 50 **then**
                $\mathcal{C}_{\text{synthetic}} \leftarrow \mathcal{C}_{\text{synthetic}} \cup \{(\mathcal{R}_{e,r}, r)\}$
            end if
        end for
    end for
    $\mathcal{C}_{\text{synthetic}} \leftarrow \underset{\substack{\mathcal{C}'_{\text{synthetic}} \subset \mathcal{C}_{\text{synthetic}}, \\ |\mathcal{C}'_{\text{sythetic}}| = K}}{\operatorname{argmax}} \sum_{(\mathcal{R}'_{e,r}, r) \in \mathcal{C}'_{\text{synthetic}}} \sum_{i \in \mathcal{R}'_{e,r}} \text{sim}(r, r_i))$
    if $T = \texttt{contrastive}$ **then**
        return $\mathcal{C}_{\text{synthetic}}$
    else if $T = \texttt{common}$ **then**
        sampling counterpart entity's reviews $\mathcal{R}^{\text{CP}}_{e',r'}$
        $\mathcal{C}'_{\text{synthetic}} \leftarrow \{\}$
        for all $(\mathcal{R}_{e,r}, r) \in \mathcal{C}_{\text{synthetic}}$ **do**
            $(\mathcal{R}^{\text{CP}}_{e',r'}, r') \leftarrow \underset{(\mathcal{R}_{e',r'}, r') \in \mathcal{C}_{\text{synthetic}} \setminus \{(\mathcal{R}_{e,r}, r)\}}{\operatorname{argmax}} \text{sim}(r, r')$
            $\mathcal{C}'_{\text{synthetic}} \leftarrow \mathcal{C}'_{\text{synthetic}} \cup \{(\mathcal{R}_{e,r}, \mathcal{R}^{\text{CP}}_{e',r'}, r)\}$
        end for
        return $\mathcal{C}'_{\text{synthetic}}$
    end if
end procedure

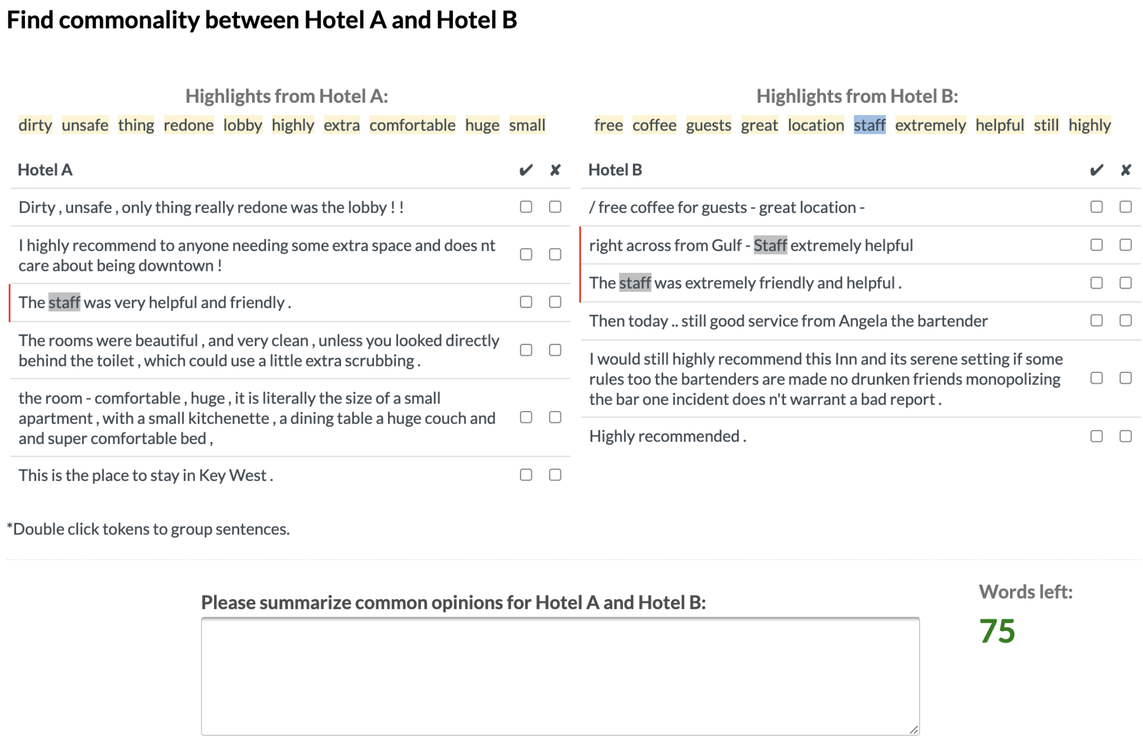

Figure 6: Summary Collection Task. We show workers two group of sentences based on labels we collected from the sentence annotation task. Similar features, such as allowing workers to group sentences that contain the same token through double clicking, are also supported in this task.

Contrastive	Summarization F1↑				Intra-Distinctiveness F1↓			
	R1	R2	RL	BS	R1	R2	RL	BS
Self-supervised								
Original (Eq. (1))	**40.78**	**10.66**	**21.53**	**29.90**	**43.89**	**17.67**	**29.13**	**34.90**
$p_{\text{cont}}^{B} \rightarrow p_{\text{comm}}^{A \cap B}$	40.60	10.50	21.36	29.69	46.95	18.03	30.00	39.73
Mixture-of-Experts	3.14	0.35	3.02	-48.33	100.00	100.00	100.00	100.00
Few-shot								
Original (Eq. (1))	42.22	**12.11**	**24.13**	**35.63**	**35.02**	8.39	21.74	28.23
$p_{\text{cont}}^{B} \rightarrow p_{\text{comm}}^{A \cap B}$	**42.35**	11.52	23.58	34.51	36.19	**7.96**	**21.03**	**27.08**
Mixture-of-Experts	3.14	0.35	3.02	-48.33	100.00	100.00	100.00	100.00

Common	Summarization F1↑				Inter-Distinctiveness F1↓			
	R1	R2	RL	BS	R1	R2	RL	BS
Self-supervised								
Original (Eq. (2))	**41.13**	**15.25**	**30.60**	**54.65**	**50.28**	**30.12**	**44.46**	**59.81**
Product-of-Experts	39.68	14.36	28.52	52.91	57.15	36.10	48.53	61.43
Few-shot								
Original (Eq. (2))	**46.80**	**20.68**	**35.62**	**61.52**	**65.14**	**43.03**	**55.15**	**70.24**
Product-of-Experts	44.68	18.32	34.18	59.76	70.11	52.23	67.61	76.26

Table 7: Summarization performance and Intra/Inter-Distinctiveness scores by CoCoSum with different Co-decoding configurations.

Contrastive	Summarization F1↑				Intra-Distinctiveness F1↓			
	R1	R2	RL	BS	R1	R2	RL	BS
SELF-THEN-FEW	**43.65**	**12.83**	**24.93**	**35.42**	**39.63**	**11.80**	**25.28**	**30.72**
MULTI-TASKING	40.81	11.37	22.25	30.27	52.98	26.40	35.65	43.28
ONLY FEW-SHOT	43.10	12.44	23.99	33.28	42.65	14.65	27.05	29.82
Common	Summarization F1↑				Inter-Distinctiveness F1↓			
	R1	R2	RL	BS	R1	R2	RL	BS
SELF-THEN-FEW	**45.90**	19.59	**34.40**	**59.32**	53.87	**29.08**	**37.94**	60.96
MULTI-TASKING	44.64	17.36	33.87	58.37	**53.13**	29.87	42.45	**59.12**
ONLY FEW-SHOT	42.56	**20.07**	32.11	57.87	62.08	44.14	49.59	64.30

Table 8: Comparisons of different few-shot learning strategies for contrastive and common opinion summarization. SELF-THEN-FEW further fine-tunes the self-supervised models using few-shot training data; MULTITASKING trains base summarization models with the pseudo review-summary data (used for self-supervised models) and few-shot training data jointly; ONLY FEW-SHOT fine-tunes a pre-trained model (i.e., `led-base-16384` in this paper) only using few-shot training data.

Computing infrastructure	NVIDIA A100
Training duration	Self-supervision: 12 hours, Few-shot learning: 1 hours
Search strategy	Manual tuning
Model implementation	`https://github.com/megagonlabs/cocosum`
Model checkpoint - self supervised	`https://huggingface.co/megagonlabs/cocosum-cont-self`
Model checkpoint - few-shot	`https://huggingface.co/megagonlabs/cocosum-cont-few`

Hyperparameter	**Search space**	**Best assignment**
# of training data for self-supervision	*choice*[25k, 50k, 100k, 200k]	50k
# of training steps for self-supervision	50,000	50,000
validation interval for self-supervision	5,000	5,000
Few-shot learning strategy	*choice*[SELF-THEN-FEW, MULTI-TASKING, ONLY FEW-SHOT]	SELF-THEN-FEW
# of training steps for few-shot learning	1,000	1,000
validation interval for few-shot learning	100	100
batch size	8	8
initial checkpoint	`allenai/led-base-16384`	`allenai/led-base-16384`
label-smoothing (Szegedy et al., 2016)	*choice*[0.0, 0.1]	0.1
learning rate scheduler	linear schedule with warmup	linear schedule with warmup
warmup steps for self-supervision	1000	1000
warmup steps for few-shot learning	100	100
learning rate optimizer	AdamW (Loshchilov and Hutter, 2019)	AdamW (Loshchilov and Hutter, 2019)
AdamW β_1	0.9	0.9
AdamW β_2	0.999	0.999
learning rate	*choice*[1e-5, 1e-4, 1e-3]	1e-5
weight decay	*choice*[0.0, 1e-3, 1e-2]	1e-3
gradient clip	1.0	1.0

Table 9: COCOSUM search space and the best assignments for contrastive opinion summarization on COCOTRIP dataset.

Computing infrastructure	NVIDIA A100
Training duration	Self-supervision: 2 hours, Few-shot learning: 30 minutes
Search strategy	Manual tuning
Model implementation	`https://github.com/megagonlabs/cocosum`
Model checkpoint - self supervised	`https://huggingface.co/megagonlabs/cocosum-comm-self`
Model checkpoint - few-shot	`https://huggingface.co/megagonlabs/cocosum-comm-few`

Hyperparameter	**Search space**	**Best assignment**
# of training data for self-supervision	*choice*[1k, 5k, 10k, 20k]	5k
# of training steps for self-supervision	5,000	5,000
validation interval for self-supervision	500	500
Few-shot learning strategy	*choice*[SELF-THEN-FEW, MULTI-TASKING, ONLY FEW-SHOT]	SELF-THEN-FEW
# of training steps for few-shot learning	1000	1000
validation interval for few-shot learning	100	100
batch size	8	8
initial checkpoint	`allenai/led-base-16384`	`allenai/led-base-16384`
label-smoothing (Szegedy et al., 2016)	*choice*[0.0, 0.1]	0.1
learning rate scheduler	linear schedule with warmup	linear schedule with warmup
warmup steps for self-supervision	1000	1000
warmup steps for few-shot learning	100	100
learning rate optimizer	AdamW (Loshchilov and Hutter, 2019)	AdamW (Loshchilov and Hutter, 2019)
AdamW β_1	0.9	0.9
AdamW β_2	0.999	0.999
learning rate	*choice*[1e-5, 1e-4, 1e-3]	1e-5
weight decay	*choice*[0.0, 1e-3, 1e-2]	1e-3
gradient clip	1.0	1.0

Table 10: COCOSUM search space and the best assignments for common opinion summarization on COCOTRIP dataset.

COCOSUM	
Entity ID: 482693 & 1547281 The staff at the hotel were very helpful and friendly. **The hotel is in a great location and close to the canal.**	**Entity ID: 202988 & 233491** The staff at the hotel were very friendly and helpful. **The hotel is ideally located for a stay in Florence.**
COCOSUMw/o Co-decoding	
Entity ID: 482693 & 1547281 The staff at the hotel are very friendly and the hotel is recommended.	**Entity ID: 202988 & 233491** This hotel is in an excellent location and the staff are very friendly and helpful.

Table 11: Common summaries generated by COCOSUM with and w/o Co-decoding for two example entity pairs. Entity-pair specific (common) opinions are highlighted in **green** (magenta).