Comparative Opinion Summarization via Collaborative Decoding

Anonymous ACL submission

Abstract

Opinion summarization focuses on generating summaries that reflect popular opinions of multiple reviews for a single entity (e.g., a hotel or a product.) While generated summaries offer general and concise information about a particular entity, the information may be insufficient to help the user compare multiple entities. 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 given sets of reviews of different entities. We develop a comparative summarization framework CoCoSum, which consists of two few-shot 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.

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 entities based on initial conditions (e.g., area, price range, restaurant type), the user often has to compare a few entities in depth by carefully reading the reviews to make a final decision (Payne et al., 1991). However, it is time-consuming and difficult for the user to detect differences and similarities between the entities, 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; Iso et al., 2021), whose goal is to generate a summary that represents salient opinions in input reviews. 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 distinctive 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

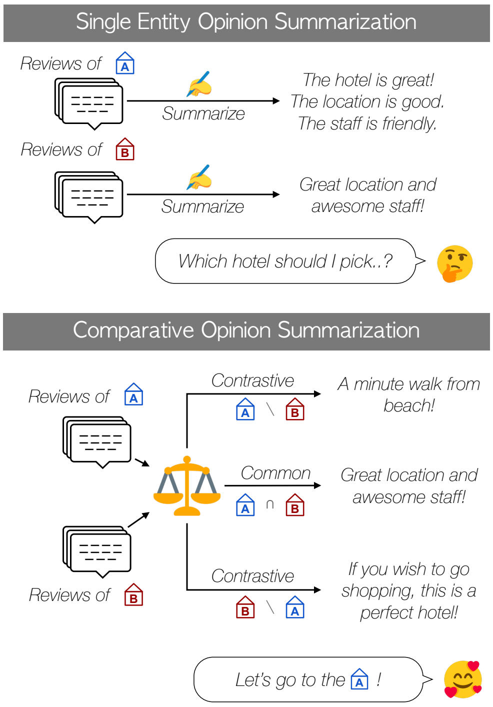
Single Entity Opinion Summarization
Reviews of A
Summarize
The hotel is great! The location is good. The staff is friendly.
Reviews of B
Summarize
Great location and awesome staff!
Which hotel should I pick..?
Comparative Opinion Summarization
Reviews of A
Contrastive
A \ B
A minute walk from beach!
Common
A ∩ B
Great location and awesome staff!
Reviews of B
Contrastive
B \ A
If you wish to go shopping, this is a perfect hotel!
Let's go to the A !

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.

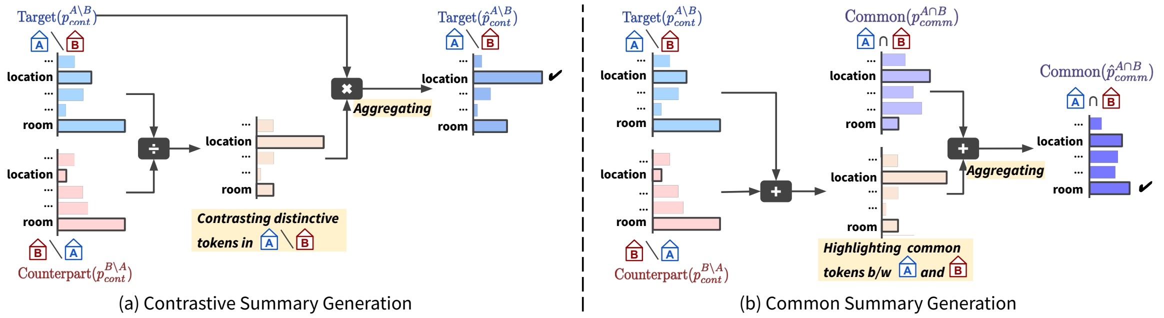
Target($p_{cont}^{A\setminus B}$)
A \ B
location
room
Counterpart($p_{cont}^{B\setminus A}$)
B \ A
÷
location
room
Contrasting distinctive tokens in A \ B
×
Aggregating
Target($\hat{p}_{cont}^{A\setminus B}$)
location
room
(a) Contrastive Summary Generation
Target($p_{cont}^{A\setminus B}$)
location
room
Counterpart($p_{cont}^{B\setminus A}$)
+
Common($p_{comm}^{A\cap B}$)
A ∩ B
location
room
Highlighting common tokens b/w A and B
+
Aggregating
Common($\hat{p}_{comm}^{A\cap B}$)
location
room
(b) Common Summary Generation

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.

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 jointly uses the two models for contrastive and common summary generation. The main idea of Co-decoding is to jointly use two summarization models by aggregating the token probability distributions in the decoding step, so the models can generate more distinctive and entity-pair-specific summaries.

Experimental results on a newly created 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 50 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 a *contrastive summary* of a target entity A against a counterpart entity B $y_{\text{cont}}^{A\setminus B}$ as a summary that describes salient opinions in $\mathcal{R}_A$ but not in $\mathcal{R}_B$. Similarly, we define a *common summary* $y_{\text{comm}}^{A\cap B}$ of entities A and B as a summary that describes common opinions in $\mathcal{R}_A$ and $\mathcal{R}_B$. Note that $y_{\text{comm}}^{A\cap B}$ and $y_{\text{comm}}^{B\cap A}$ 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.

	Task	# of Ent	Inp. Review	# of Summ.	Inp. len	Summ. len	Domain
CoCoTrip (This work)	Contrastive Common	100 50	16	300 150	1529.4	132.9 20.3	Hotels
Bražinskas et al. (2020a)	Single	100	8	300	481.3	61.2	Businesses
Bražinskas et al. (2020a)	Single	60	8	180	469.6	59.6	Products
Chu and Liu (2019)	Single	200	8	200	581.1	70.4	Businesses
Bražinskas et al. (2020b)	Single	60	8	180	473.4	59.8	Products

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.

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 50 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. Our dataset contains a similar scale of summaries to existing abstractive opinion summarization datasets, and the input reviews are about three times longer than others.

3 CoCoSum

For single-entity opinion summarization, input reviews can be used as pseudo summaries for training summarization models in a self-supervised fashion. This approach is not suitable for comparative opinion summarization as the task takes two sets of reviews for different entities to generate contrastive and common summaries, which have significantly different characteristics from the original review as supported by Table 1. In addition, recent studies have shown the effectiveness of pre-trained Transformer models for summarization tasks (Zhang et al., 2020; Oved and Levy, 2021).

Therefore, we use a few-shot learning approach that fine-tunes a pre-trained Transformer model using input reviews and corresponding reference summaries. However, while the few-shot learning approach helps the model acquire the writing style, we found that it was not sufficient to learn to generate summaries that contain distinctive and common opinions between two entities. This led us to design a “collaborative” decoding solution Co-decoding, which calculates the token probability distribution based on two summarization models trained for common and contrastive summary generation.

3.1 Base Summarization Model

CoCoSum consists of two summarization models that are separately fine-tuned using reference contrastive and common summaries, respectively. Both summarization models take concatenated reviews of two entities as input. To distinguish which reviews are about which entity, we introduce additional *type embeddings* into the input layer of the encoder to distinguish which reviews are about the target or counterpart entity, as shown in Figure 3.

For contrastive summary generation (i.e., $y_{\text{cont}}^{A\setminus B} \neq y_{\text{cont}}^{B\setminus A}$), we keep the original order of the target entity and counterpart entity as the model should recognize which one is the target entity. Then, we fine-tune a pre-trained Transformer model using reference summaries for entity pairs.

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 augment training data by creating both concatenation orders for fine-tuning. For the inference time, 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.

We refer to the base summarization model for contrastive (common) summary generation as the *contrastive (common) summarization model.*

3.2 Collaborative Decoding

Although few-shot learning is an effective solution for training summarization models, the model may

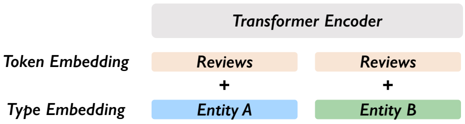

Figure 3: Encoder of the base summarization model has *type embeddings* to distinguish the original entity.

not be sufficient to generate contrastive and common summaries. This is because such models do not have the functionality to *compare and contrast* two summarization models for better contrastive and common summary generation. To incorporate direct interactions between models, we design a solution Co-decoding that uses two summarization models in the decoding phase, which would help generate better contrastive and common summaries than individual models as illustrated in Figure 2.

We denote the token probability distribution of a model $M \in \{\text{cont}, \text{comm}\}$ at t-th step by $P_M(Y_t \mid y_{<t}, \mathcal{R}_A, \mathcal{R}_B)$. The key idea of Co-decoding is to aggregate $P_{\text{cont}}(\cdot)$ and $P_{\text{comm}}(\cdot)$ 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 counterpart 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:

$$\hat{p}^{A\setminus B}_{\text{cont}}(Y_t) \propto p^{A\setminus B}_{\text{cont}}(Y_t)\left(\frac{p^{A\setminus B}_{\text{cont}}(Y_t)}{p^{B\setminus A}_{\text{cont}}(Y_t)}\right)^{\delta}, \tag{1}$$

where $p^{A\setminus B}_{\text{cont}}(Y_t) := P_{\text{cont}}(Y_t \mid y_{<t}, \mathcal{R}_A, \mathcal{R}_B)$ is the token probability for a contrastive summary $\hat{y}^{A\setminus B}_{\text{cont}}$. Note that for both $p^{A\setminus B}_{\text{cont}}(Y_t)$ and $p^{B\setminus A}_{\text{cont}}(Y_t)$, 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 $\frac{p^{A\setminus B}_{\text{cont}}(Y_t)}{p^{B\setminus A}_{\text{cont}}(Y_t)}$ (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:

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

where $p^{A\cap B}_{\text{comm}}(Y_t) := P_{\text{comm}}(Y_t \mid y_{<t}, \mathcal{R}_A, \mathcal{R}_B)$ is the token probability distribution of the common summary model.

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: $p^{A\setminus B}_{\text{cont}}(Y_t) + p^{B\setminus A}_{\text{cont}}(Y_t)$ (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 used CoCoTrip for the evaluation. For robust evaluation, we ran the training and evaluation

	Contrastive			Common			Pair
	R1 ↑	R2 ↑	RL ↑	R1 ↑	R2 ↑	RL ↑	DS ↑
Unsupervised Extaractive							
LexRank (Erkan and Radev, 2004)	23.28	3.68	13.85	21.82	4.17	14.50	43.69
LexRank$_{\text{BERT}}$ (Reimers and Gurevych, 2019)	27.64	5.31	15.89	22.38	4.54	15.44	40.51
Unsupervised Abstractive							
MeanSum (Chu and Liu, 2019)	33.72	7.83	19.61	13.77	0.98	10.56	70.04
OpinionDigest (Suhara et al., 2020)	37.27	8.91	20.77	21.01	4.02	14.87	72.94
CopyCat (Bražinskas et al., 2020b)	23.19	6.43	16.23	35.35	11.55	24.05	39.34
BiMeanVAE (Iso et al., 2021)	37.87	9.82	22.20	37.07	14.17	26.39	40.59
CoCoSum	39.05	10.17	21.51	39.38	15.06	30.11	**80.02**
w/o Co-decoding	40.96	11.19	23.15	40.36	16.14	31.48	74.40
Human upper bound	47.29	12.75	26.15	49.11	18.25	37.76	78.59

Table 2: ROUGE 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.

process 5 times with different train/dev/test splits (40%/20%/40%) and report the average scores.

For both contrastive and common summarization models, we fine-tuned 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] We used Adam optimizer (Kingma and Ba, 2015) with a linear scheduler with an initial learning rate of 0.002 and a warm-up step of 1000. 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 $p^{A\setminus B}_{\text{cont}}(Y_t)$, $p^{B\setminus A}_{\text{cont}}(Y_t)$, 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.

We compare CoCoSum with a variety of opinion summarization models as baselines, namely LexRank (Erkan and Radev, 2004), MeanSum (Chu and Liu, 2019), OpinionDigest (Suhara et al., 2020), CopyCat (Bražinskas et al., 2020b), and BiMeanVAE (Iso et al., 2021).

4.2 Automatic Evaluation

Evaluation Metrics For summarization quality, we use ROUGE 1/2/L F1 scores (Lin, 2004)[2] 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:

$$\text{DS} = 1 - \frac{\sum_{(i,j)\in I^{(2)}} |V_i \cap V_j| - 2|\bigcap_{i\in I} V_i|}{|\bigcup_{i\in I} V_i|},$$

where $I = \{A, B, C\}$, $I^{(2)}$ is the 2-subsets of I and V_A, V_B, V_C denote the token sets of two generated contrastive summaries $\hat{y}^{A\setminus B}_{\text{cont}}$, $\hat{y}^{B\setminus A}_{\text{cont}}$, and a generated common summary $\hat{y}^{A\cap B}_{\text{comm}}$, respectively.

Results As shown in Table 2, CoCoSum outperforms the baseline methods for the ROUGE scores (summarization quality) and the distinctiveness score (DS), showing the effectiveness of few-shot learning and Co-decoding. Comparing the ROUGE scores by CoCoSum and CoCoSum w/o Co-decoding, we confirm that Co-decoding sacrifices the summarization performance as expected while significantly improving the distinctiveness, achieving the same quality level as the gold-standard summaries.

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

First, we show human annotators four summaries, including three summaries generated by CoCoSum, CoCoSum w/o Co-decoding, and BiMeanVAE, and one human-written summary. We then ask annotators to select the best and worst summary according to three different criteria, i.e., informativeness, coherence, and non-redundancy. We then calculate the scores using best-worst scaling (Louviere et al., 2015) with values ranging from -1.0

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

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

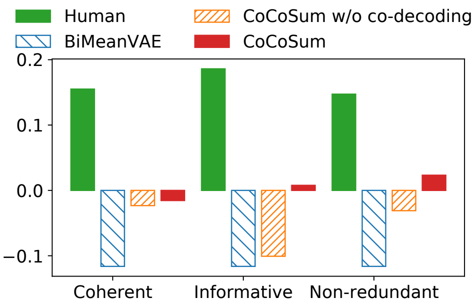

Figure 4: Human evaluation with best-worst scaling.

	Not	Partial	Overlap
BiMeanVAE	18.5%	21.3%	60.2%
CoCoSum	**76.4%**	**9.0%**	**14.6%**
w/o Co-decoding	58.0%	22.0%	20.0%

Table 3: Human evaluation on content overlap

(unanimously worst) to +1.0 (unanimously best). As shown in Figure 4, human-written summaries show much better performance than the automatically generated summaries. While among the automatically generated summaries, CoCoSum shows better performance on all three criteria.

Second, we ask 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 ask 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. This result also aligns with our automatic evaluation on the distinctiveness (Table 2), and it demonstrates that CoCoSum can produce more distinctive contrastive and common summaries.

Lastly, we conduct a summary content support study to evaluate how faithful the generated summaries are toward the input reviews. The results indicate that all methods show comparable performance while CoCoSum is slightly better than the others. The results are presented in the Appendix.

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.

Contrastive (Intra-ROUGE F1↓)	R1	R2	RL
BiMeanVAE (Iso et al., 2021)	68.23	49.12	54.81
CoCoSum	**32.75**	**7.39**	**18.98**
w/o Co-decoding	41.54	14.54	26.93
Human upper bound	38.07	7.94	20.17
Common (Inter-ROUGE F1↓)	R1	R2	RL
BiMeanVAE (Iso et al., 2021)	71.61	50.52	59.84
CoCoSum	**55.69**	**37.93**	**50.35**
w/o Co-decoding	82.31	70.91	78.54
Human upper bound	38.18	16.72	30.11

Table 4: Intra-ROUGE scores for contrastive summary generation (above) and Inter-ROUGE scores for common summary generation (below.)

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 ROUGE (Intra-ROUGE) scores* only between two contrastive summaries for each entity pair to measure the *intra-entity-pair distinctiveness*.

Table 4 (above) shows that CoCoSum significantly outperforms the state-of-the-art opinion summarization model (BiMeanVAE) and the ablated version of CoCoSum (i.e., w/o Co-decoding.) The results confirm that Co-decoding successfully generates contrastive summaries that contain distinctive opinions of each other.

Does Co-decoding address the overly generic summary issue for common summaries? While the base few-shot learning summarization model suffers from generating overly generic summaries, CoCoSum with Co-decoding should alleviate this issue since it highlights entity-pair specific tokens from the contrastive summarization model. To verify this, we use an alternative distinctiveness metric—the *inter-entity-pair ROUGE (Inter-ROUGE) scores*.

Similar to the Intra-ROUGE scores, CoCoSum also shows strong performance for the Inter-ROUGE scores as shown in Table 4 (below.) The results confirm that Co-decoding successfully addresses the overly generic summary issue, indicating that CoCoSum generates a meaningful common summary for each entity pair.

CoCoSum	*The hotel was available with a deal via the hotel* **, but there were some issues with the elevator and lines were a bit plain. Overall this is a perfect hotel for solo stays in Rome and not far from Campiano Airport. The rooms in the hotel are not huge but comfortable** and clean. **The bathrooms are gorgeous and the rooms make the day extra special. The hotel upgraded rooms to have Boschari toiletries on the bed each day. The elevator was a bit plain** *and the lines were too lines.* The hotel staff are always courteous and helpful. **Every member of staff have loads of great advice and recommendations for local attractions and sight-seeing. The hotel provides a good size buffet and on roof top garden** *you can enjoy a nice shower.*
w/o Co-decoding	This is a perfect hotel for any type of stay and you will want to keep coming back for the tranquillity, *unbeatable price* and the great service. This hotel is in a really bustling area of Rome and close to the main sights of the city. **The rooms in the hotel are a good size, with spacious bathrooms and even some really great chocolates on the bed.** The hotel staff are very helpful and always willing to help out with their polite manners. The breakfast provided by the hotel was really good, *although a little bit basic.* **The elevator in this hotel is a little bit old but it's in good condition.**

Table 5: Contrastive summaries (Entity ID: 203083) generated by CoCoSum with and w/o Co-decoding. **Distinctive (desired)** / common (undesired) opinions are color-coded and *hallucinated content* is in *italics*.

	Entity IDs: 203083 & 208552	**Entity IDs: 305947 & 305813**
CoCoSum	The staff at the hotel were very helpful and friendly. **The hotel is situated in a bustling area of Rome.**	The staff at the hotel were very helpful and nice to guests. **The hotel is in a working class area of Kowloon.**
w/o Co-decoding	The staff at the hotel are friendly and the rooms are clean.	The staff at the hotel are very helpful and the rooms are clean.

Table 6: Common summaries generated by CoCoSum with and w/o Co-decoding for two example entity pairs. Common opinions are in magenta and **entity-pair specific** opinions are highlighted in **bold**.

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.

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: $p^{A\setminus B}_{\text{cont}}(Y_t) + \left(p^{A\setminus B}_{\text{cont}}(Y_t)/p^{B\setminus A}_{\text{cont}}(Y_t)\right)^{\delta}$.

With this configuration, we observe significant degradation of summarization quality (e.g., 11.05 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 $p^{A\setminus B}_{\text{cont}}(Y_t)/p^{A\cap B}_{\text{comm}}(Y_t)$. It shows competitive performance as the original design with respect to the Intra-ROUGE scores (e.g., 33.13 on Intra-R1). However, this configuration does not perform well in the summarization performance (e.g., 34.90 on R1.) This may be attributed to the fact that the contrastive and common summaries have significantly different characteristics, especially in the writing style and the summary length. Therefore, when the decoding step goes beyond the average length of common summaries, the common summary generation model might not provide a meaningful token probability distribution, which can harm summary generation by Co-decoding.

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: $p^{A\cap B}_{\text{comm}}(Y_t)\prod_{E\in\{A\setminus B, B\setminus A\}} p^{E}_{\text{cont}}(Y_t)^{\gamma}$.

This configuration performs competitively with the original Co-decoding for the standard ROUGE scores while the Inter-ROUGE scores were significantly degraded (e.g., 39.86 on R1, 61.28 on Inter-R1.) 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

Contrastive Summary Generation Table 5 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 contains more contrastive opinions for users to compare the entities.

Common Opinion Summarization Table 6 shows examples of common summaries generated by CoCoSum with and w/o Co-decoding for two entity pairs. Compared to w/o method, CoCoSum can generate common summaries that contain entity-pair specific opinions in addition to common opinions. Meanwhile, CoCoSum w/o Co-decoding generates summaries with generic opinions, which is a limitation of the few-shot learning approach as it is biased by the training data.

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 et al., 2021; Elsahar et al., 2021; Im et al., 2021; Wang and Wan, 2021; Isonuma et al., 2021, *inter alia*).

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 been also 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 few-shot 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.

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.

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.

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.

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.

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.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR 2015*.

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.

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

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.

J Payne, JR Bettman, and EJ Johnson. 1991. Consumer decision making. *Handbook of consumer behaviour*, pages 50–84.

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.

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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

	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 7: 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.

Contrastive	ROUGE F1↑			Intra-ROUGE F1↓		
	R1	R2	RL	R1	R2	RL
Original (Eq. (1))	**39.05**	**10.17**	**21.51**	32.75	7.39	18.98
Mixture-of-Experts	11.05	1.12	6.93	*3.56*	*0.04*	*2.51*
$p_{\text{cont}}/p_{\text{comm}}$	34.90	7.34	18.32	33.13	7.42	18.32

Common	ROUGE F1↑			Inter-ROUGE F1↓		
	R1	R2	RL	R1	R2	RL
Original (Eq. (2))	39.38	15.06	30.11	**55.69**	**37.93**	**50.35**
Product-of-Experts	**39.86**	**15.08**	**30.59**	61.28	43.44	55.27

Table 8: Summarization performance and Intra/Inter-ROUGE scores by CoCoSum with different Co-decoding configurations. The mixture-of-experts configuration does not generate contrastive summaries with an acceptable quality. Thus, its low Intra-ROUGE scores are not meaningful.

A Comparative Opinion Summarization

Table 7 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 TripAdviros 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.

We collected 3 annotations 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.

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. For every entity pair, we collected 3 reference summaries for each of *two* contrastive summary generation and *one* common summary generation tasks.

	Full (↑)	Partial (↑)	No (↓)
BiMeanVAE	60.9%	24.0%	15.1%
CoCoSum	**62.2%**	**24.6%**	**13.1%**
w/o Co-decoding	60.6%	26.1%	13.3%

Table 9: Human evaluation on content support

C Additional Evaluation Results

C.1 Baselines

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

LexRank (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[3];

MeanSum (Chu and Liu, 2019): the unsupervised single entity opinin summarization solution[4];

OpinionDigest (Suhara et al., 2020): a weakly supervised opinion summarization approach[5];

CopyCat (Bražinskas et al., 2020b): a single entity opinion summarization solution based on leave-one-out reconstruction[6];

BiMeanVAE (Iso et al., 2021): an optimized single entity opinion summarization solution[7] for MeanSum.

C.2 Human Evaluation

Table 9 shows the human evaluation results on content support. For every sentence in the generated contrastive/common summary, we obtain the review sentences that human annotators labeled as "contrastive"/"common" during the CoCoTrip creation process, and ask human annotators to judge if the summary sentence is fully, partially, or not supported by the corresponding reviews. As shown, all methods show compatible performance while CoCoSum is slightly better than the others.

D Full qualitative example

Table 10 shows the full generated contrastive summaries shown in Table 5 for the entity pair, 203083 and 208552.

[3] https://github.com/UKPLab/sentence-transformers

[4] https://github.com/sosuperic/MeanSum

[5] https://github.com/megagonlabs/opiniondigest

[6] https://github.com/abrazinskas/Copycat-abstractive-opinion-summarizer

[7] https://github.com/megagonlabs/coop

CoCoSum	Intra-ROUGE1/2/L = (**41.48, 5.97, 18.52**)
Entity ID: 203083 *The hotel was available with a deal via the hotel* **, but there were some issues with the elevator and lines were a bit plain . Overall this is a perfect hotel for solo stays in Rome and not far from Campiano Airport . The rooms in the hotel are not huge but comfortable** and clean . **The bathrooms are gorgeous and the rooms make the day extra special . The hotel upgraded rooms to have Boschari toiletries on the bed each day . The elevator was a bit plain** *and the lines were too lines* . The hotel staff are always courteous and helpful . **Every member of staff have loads of great advice and recommendations for local attractions and sight-seeing . The hotel provides a good size buffet and on roof top garden** *you can enjoy a nice shower* .	**Entity ID: 208552** **Hotel Campo de Fiori is great for sight-seeing in Rome** *and is n't cheap* **because it 's European-sized** but it 's beautiful and well appointed . The hotel is situated near the ancient centre of Rome **but a 5 minute walk to most of the sites , 20-minute walk to the library and restaurant La Scalla and The Library are both great restaurants but the rooms here are smaller in size but they are decorated with character and have a great view of the historic area .** The staff at the hotel were extremely helpful and made you feel so welcome there . The food is good here and offers both breakfast **and orange juice . It 's a little disappointing to see the food restrict to the basics of the food as it is in a claustrophobic room .**
CoCoSum w/o Co-decoding	Intra-ROUGE1/2/L = (59.14, 26.67, 35.02)
Entity ID: 203083 This is a perfect hotel for any type of stay and you will want to keep coming back for the tranquillity , *unbeatable price* and the great service . This hotel is in a really bustling area of Rome and close to the main sights of the city . **The rooms in the hotel are a good size , with spacious bathrooms and even some really great chocolates on the bed .** The hotel staff are very helpful and always willing to help out with their polite manners . The breakfast provided by the hotel was really good , *although a little bit basic* . **The elevator in this hotel is a little bit old but it 's in good condition .**	**Entity ID: 208552** This is a perfect hotel with a central location and **superb roofterrace . It is also a great place to drink some wine in the hours leading up to sunset .** The location of this hotel is great **for those wanting to relax** and have a few drinks , **as well as the view deck in the room was really great . The hotel 's rooms are smaller in size** but are clean **and with a good view of the historic area of Rome .** The bathroom in the room was great . The hotel staff are so helpful and always willing to help out with their polite manners . The hotel provides a great breakfast and both the breakfast **and the wine will be great . The hotel 's location is excellent for those wanting to relax** and have a few drinks , *as there is a great view over the river* .

Table 10: Contrastive summaries generated by CoCoSum with and w/o Co-decoding for an example entity pair. Distinctive (common) opinions are highlighted in **blue** (orange), and hallucinated content is in italics.