Generating Category/Aspect-based Review Summaries Using ACOS Quadruples

Pooya Adami
padami@usc.edu

Xinyue Cui
xinyuecu@usc.edu

Victor Hui
victoryi@usc.edu

Saeedeh Mahmoodifar
saeedehm@usc.edu

Pinyi Wang
pinyiw@usc.edu

Abstract

This project aims to explore various techniques for generating category or aspect-based summaries across multiple restaurant reviews using Aspect-Category-Opinion-Sentiment (ACOS) quadruples. We utilized OpenAI's GPT-3.5 model for ACOS labeling and evaluated its performance on Laptop-ACOS, Restaurant-ACOS, and Yelp datasets. We also experimented with different grouping methods and GPT prompts for generating summaries and employed novel evaluation metrics to assess the quality of the generated summaries. The results showed promising outcomes, suggesting that data augmentation using ACOS quadruples could be a promising approach for improving the quality of aspect-based summarization. The outcomes of this study can have a positive impact on review sites, enabling consumers to make more informed decisions and ultimately enhancing their overall satisfaction. Further research can explore other techniques for generating summaries based on ACOS quadruples and evaluation metrics to improve the quality of aspect-based summarization.

1 Introduction

In the field of Aspect-Based Sentiment Analysis (ABSA), recent advances have exhibited potential in extracting Aspect-Category-Opinion-Sentiment (ACOS) quadruples from opinionated content, such as online reviews. However, most existing studies have focused on extracting opinions and sentiment from one review at a time. This project aims to investigate different techniques for generating summaries over multiple reviews of a product/restaurant/service using ACOS quadruples, and multiple evaluation metrics are employed to measure the quality of the summaries.

The project team believes that the outcomes of this study can have a positive impact on review sites. Our proposed review summarization system aims to offer context to ratings, enabling consumers to make more informed decisions and ultimately enhancing their overall satisfaction. To attain this objective, the study employs state-of-the-art ACOS quadruple extraction techniques using the fine-tuned T5 model presented in (Peper and Wang, 2022). Reviews with the same category/aspect in their ACOS quadruples are grouped together to provide an aspect-oriented review summary. In order to preserve the most crucial concepts (ACOS quadruples) during the summarization process, one method is to augment the original reviews with ACOS quadruples.

The effectiveness of the proposed model is assessed using metrics presented in (Bhaskar et al., 2022) to evaluate if the model is generating accurate (faithfulness), specific (genericity), and factual (factuality) summaries.

2 Related Work

Aspect-Based Sentiment Analysis (ABSA) is a subfield of Natural Language Processing (NLP) that focuses on identifying and extracting sentiments related to specific attributes or aspects of a product or service (Hu and Liu, 2004). Generative models have shown impressive results in ABSA tasks, such as ACOS extraction. Conditional Variational Autoencoder (CVAE) and Transformer-based models, such as BERT and GPT-2, have been used for ABSA tasks, including ACOS extraction (Li et al., 2018). Variational Autoencoder (VAE) is another generative model used to generate informative reviews by conditioning on the input text and the aspect to be reviewed. Graph Convolutional Network (GCN) is effective in capturing relationships between aspects, categories, opinions, and sentiments in the text (Li et al., 2020).

Common ABSA subtasks include recognizing and categorizing aspect terms, identifying supporting opinion terms, and detecting sentiment polarity expressed in text. However, existing models struggle with implicit language in over 30% of sentiment expressions (Cai et al., 2021). Implicit aspects and opinions are particularly difficult to recognize; therefore, addressing the ACOS quadruple extraction task is crucial, especially for formulations supporting implicit aspects and opinions.

Generative models have been introduced with techniques for improved structured generation for ACOS quadruple extraction. For example, Peper and colleagues proposed a model to improve quadruple prediction by using supervised contrastive learning. However, these models focus on extracting ACOS quadruples on one review at a time (Peper and Wang, 2022). This project aims to explore various methods for generating summaries of multiple reviews based on ACOS quadruples resulting from the improved models. Additionally, the study investigates evaluation metrics to measure the quality of review summaries.

3 Problem Description

Contemporary review platforms such as Yelp or Amazon typically utilize a 1-to-5-star rating system to convey the quality of their products. Our project's motivation arises from our personal experiences, that these review websites often provide inadequate information to guide users effectively. These system's fundamental flaw lies in its inability to capture the complex and subjective nature of individual opinions. Users with varying priorities and preferences will assign conflicting ratings to the same product/establishment. Providing us an opportunity to develop a system that generates concise, category-based summaries, with our goal being to create category/aspect based review summaries that better inform users, allowing them to make decisions according to their unique preferences and priorities.

We hope this project can ignite discussions that may ultimately lead to enhancements in the current review standards and potentially create a superior product for universal use.

4 Methods

Our project repository can be found here[1].

4.1 Materials

We used three primary datasets for our study: the Laptop-ACOS and Restaurant-ACOS datasets (Cai et al., 2021), along with Yelp's business review dataset (Inc., 2022). We needed the Yelp dataset because the Laptop-ACOS and Restaurant-ACOS datasets contain reviews of multiple products (laptops and restaurants, respectively) without any product IDs to group by. Our project's goal was to summarize reviews of a particular product, so this posed a significant challenge.

To obtain an ACOS labeled dataset for a product, we first evaluated the effectiveness of ACOS labeling through OpenAI's GPT-3.5[2] model, using few-shot learning techniques with the Laptop-ACOS dataset. We then used the Restaurant-ACOS dataset as few-shot examples to label Yelp's reviews with ACOS quadruples. Finally, we leveraged the labeled Yelp dataset to investigate various techniques for generating aspect-based summaries.

4.2 ACOS Labelling

We investigated two primary methods for extracting ACOS quadruples from reviews. Initially, we utilized ACOS quadruples generated by the pretrained T5 model introduced by (Peper and Wang, 2022). However, during testing, we discovered that the ACOS quadruples produced by GPT-3.5 using the few-shot learning method exhibited the highest F1 score, signifying our project's first innovation – improving upon the paper's original performance. By adopting the GPT model to generate ACOS quadruples, we eliminated the need to further experiment with fine-tuning a T5 large model for ACOS labeling, which would have been a significantly time-consuming endeavor.

Expanding on our method of generating ACOS quadruples via the use of OpenAI's GPT-3.5 model, we evaluated the performance of the model on the Laptop-ACOS dataset, since it was heuristically more challenging than the Restaurant ACOS dataset, as evidenced by their respective state-of-the-art F1 scores of 0.39 and 0.53 (Papers with Code, 2023). For picking GPT-3.5 models, although gpt-3.5-turbo is newer and cheaper, from our experience, text-davinci-003 is better for text completion tasks in general due to gpt-3.5-turbo's optimization for chat. Therefore, we used text-davinci-003 model to experiment with different prompting techniques of 5-shot and 10-shot learning.

In the original work, the "Category" element of the ACOS quadruples was drawn from a predefined set. However, to further enable the label-

[1] https://github.com/pinyiw/CSCI544-project

[2] https://platform.openai.com/docs/models/gpt-3-5

ing to work on any product type without any prior knowledge or human effort, we allowed GPT to generate and classify categories autonomously, so that it would improvise "Category" values based on the context of the reviews. For evaluation, we omitted the "Category" element when comparing with the gold quadruples.

Despite being double the cost, 10-shot learning significantly outperformed 5-shot learning. Consequently, we proceeded to use 10-shot learning to label our Yelp dataset.

N-shot	Prec.	Recall	F1	No. tokens
5	0.382	0.368	0.375	**128k**
10	**0.468**	**0.402**	**0.432**	205k

Table 1: ACOS Labeling with few-shot learning

4.3 Summary Generation

Due to resource constraints, we randomly select one retaurant in Yelp's business review dataset to perform aspect-based review summarization. We experimented with three methods of summary generation: only using the reviews; only using the ACOS quadruples; using both the reviews and their corresponding ACOS quadruples. However, when utilizing GPT to generate summaries, we encountered two challenges. The first challenge was that in some cases the summary is not concise enough. The second challenge was category leaking, which occurs when the summary of a specific group mentions topics outside of that group. For example, when generating summaries for food quality, the summary may include information about service.

We resolved these issues through prompt engineering. To address the first challenge, we instructed GPT to summarize the reviews using at most 5 sentences. For the second issue of category leaking, we requested GPT to perform summarization on the current topic only. Refining the prompts using OpenAI's prompt engineering guidelines (Shieh, 2023) also substantially enhanced both the quality of the generated summaries and the model's ability to follow the instructions. For instance, by breaking the prompts into shorter, simpler sentences, providing clear instructions on what to do rather than what not to do, and beginning with a direct task command, the prompting became much more effective.

5 Experimental Results

5.1 Evaluation Metrics

In order to assess the effectiveness of our approach, we implemented novel metrics proposed by (Bhaskar et al., 2022) that evaluates aspect-based review summaries generated by our model from three perspectives: faithfulness, factuality, and genericity. Specifically, faithfulness measures how well the summary represents the general consensus expressed in the reviews; factuality assesses how accurately the summary captures the key information in the reviews; genericity determines whether the summary is too generalized with certain words being overused.

However, as the generated summaries often contain compound sentences with contrasting opinions, it is difficult for the entailment model to accurately infer the entailment score. To address this issue, we employed the split-and-rephrase approach described in (Bhaskar et al., 2022). We prompted GPT to split the summaries into sentences with simple propositions, and 2-shot learning achieved the most satisfying results.

In terms of implementing these metrics, the faithfulness and factuality scores are derived from the SummaC entailment model (Laban et al., 2021), which computes how strongly each sentence in the split-and-rephrased summary is entailed by each piece of review, as shown in Figure 1. Specifically, we calculate faithfulness by measuring the number of reviews that entail the summary sentence with a score above a certain threshold, which is chosen to be 0.2 based on manual inspection. For factuality, we take the maximum entailment score of the summary sentence with respect to all the reviews. Finally, the genericity score is obtained by computing the (exponential) Inverse Document Frequency, given by the total number of randomly sampled restaurants (which is 10 in our experiments) divided by the number of restaurants where a word in the summary appears.

5.2 Results

We conducted experiments on different grouping methods and different GPT prompts for summarization. Tables 2 and 3 demonstrate the results for grouping by category and aspect, respectively. We can see that when grouping by category, prompting GPT with both the reviews and their corresponding ACOS quadruples produce the best results on 2 out of 3 metrics, namely Factuality and Gener-

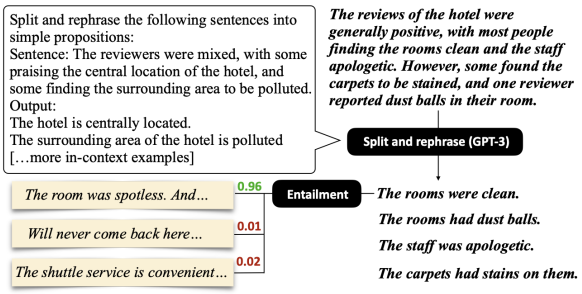
Split and rephrase the following sentences into simple propositions:
Sentence: The reviewers were mixed, with some praising the central location of the hotel, and some finding the surrounding area to be polluted.
Output:
The hotel is centrally located.
The surrounding area of the hotel is polluted
[…more in-context examples]
The reviews of the hotel were generally positive, with most people finding the rooms clean and the staff apologetic. However, some found the carpets to be stained, and one reviewer reported dust balls in their room.
Split and rephrase (GPT-3)
The room was spotless. And…
0.96
Will never come back here…
0.01
The shuttle service is convenient…
0.02
Entailment
The rooms were clean.
The rooms had dust balls.
The staff was apologetic.
The carpets had stains on them.

Figure 1: Split and rephrase the summary and compute per-sentence entailment scores. (Bhaskar et al., 2022)

GPT Prompt	Faithfulness	Factuality	Genericity
Reviews only	**27.0**	0.7091	14.543
ACOS only	22.305	0.6689	14.896
Reviews + ACOS	23.047	**0.7660**	**19.639**

Table 2: Experiment results for grouping by *category*. Best results are highlighted in bold font.

icity. This indicates that the generated summary accurately captures the essential information in the reviews and is more specific for the particular restaurant. However, when grouping by aspect, we observe that the best results for each of the three metrics are obtained from distinct prompting methods, possibly due to the wording changes in the split-and-rephrase step during evaluation that affects the summarization quality. Nevertheless, the positive outcomes we obtain with Review + ACOS prompting suggests that data augmentation using ACOS quadruples could be a promising approach for improving the quality of aspect-based summarization.

6 Conclusion

In this project, we investigated various techniques for generating category or aspect-based summaries across multiple restaurant reviews using ACOS quadruples. We utilized GPT-3.5 for ACOS labeling and evaluated its performance on Laptop-ACOS, Restaurant-ACOS, and Yelp datasets. We also explored different grouping methods and GPT prompts for generating summaries and employed novel evaluation metrics to assess the quality of the generated summaries. Our experimental results showed promising outcomes, suggesting that data augmentation using ACOS quadruples could be a promising approach for improving the quality of aspect-based summarization. We believe that our project's outcomes can have a positive impact on review sites, enabling consumers to make more informed decisions and ultimately enhancing their overall satisfaction. Further research can explore other techniques for generating summaries based on ACOS quadruples and evaluation metrics to improve the quality of aspect-based summarization.

GPT Prompt	Faithfulness	Factuality	Genericity
Reviews only	20.444	0.7752	**23.142**
ACOS only	29.531	**0.8295**	20.992
Reviews + ACOS	**32.307**	0.7803	14.654

Table 3: Experiment results for grouping by *aspect*. Best results are highlighted in bold font.

7 Future Work

We plan to examine the model's effectiveness on different restaurant datasets as part of our future endeavors. Although the split-and-rephrase step is a helpful technique in natural language processing for breaking down complex sentences into simpler ones, it can also have a negative impact on the accuracy of our summary evaluation by sometimes oversimplifying or under-simplifying the sentences and thus adding more noise to the evaluation metrics. Future work in this area should focus on developing more sophisticated algorithms and models that can better handle the split-and-rephrase process while maintaining a consistent level of simplification across summaries. This may involve exploring new techniques for identifying the most critical information in a text and rephrasing it more precisely and concisely, as well as leveraging machine learning and deep learning approaches to measure the information loss during the split-and-rephrase process. Additionally, efforts should be made to evaluate and benchmark the degree of category leaking and the quality of summarization of each aspects and categories. For more useful information to the product/service users, generating aspect-based rating scores extracted from the user ratings and sentiment of each review is also something we could work on.

8 Division of Labor

The members of the team have made significant contributions to the project. Pinyi Wang was responsible for ACOS labeling with GPT few-shot learning and evaluating summaries generated with reviews only. Victor Hui led the effort in invesitgating the GPT API and evaluating summaries generated with ACOS only, while Xinyue Cui was responsible for generating summaries using both reviews and ACOS quadruples, as well as implementing the evaluation metrics. Pooya Adami and Saeedeh Mahmoodifar conducted research on the integration of ACOS quadruples into the summarization process. These individual efforts have synergized to advance the project forward and enabled us to make substantial progress toward achieving our goals.

References

Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. 2022. Zero-shot opinion summarization with gpt-3. *ArXiv*, abs/2211.15914.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. 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 340–350.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Yelp Inc. 2022. Yelp dataset: A trove of reviews, businesses, users, tips, and check-in data!

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

P. Li, L. Huang, J. Liu, and L. Wu. 2020. A graph-based framework for aspect category opinion target extraction. In *Information Sciences*.

Y. Li, H. Chen, J. Yang, T. Bai, and X Zhang. 2018. Conditional variational autoencoder for sentiment analysis of social media text. In *IEEE Access*.

Papers with Code. 2023. Papers with code: Aspect-category-opinion-sentiment quadruple extraction leaderboard.

Joseph Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. In *Conference on Empirical Methods in Natural Language Processing*.

Jessica Shieh. 2023. Best practices for prompt engineering with openai api: How to give clear and effective instructions to gpt-3 and codex.