Status Report: Generating 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

1 Tasks performed

Our project repository can be found here at: https://github.com/pinyiw/CSCI544-project

1.1 Dataset preparation

We deep dived into the usage and evaluation of the ACOS Quadruples Extraction tasks introduced by (Cai et al., 2021) and implemented utility code to load and parse the Restaurant-ACOS and Laptop-ACOS datasets (Cai et al., 2021).

Since these two datasets include reviews of multiple products and reviewee IDs were not provided, they don't work well with the task we are solving. Hence, we are mainly using them to check the reproducibility of the fine-tuned T5 model published by the same paper (Cai et al., 2021) and further obtain our new ACOS dataset focusing on the same product from Yelp or Amazon.

1.2 ChatGPT API

We have developed foundational code and familiarized ourselves with the utilization of OpenAI's GPT API. This can be found in the "experiment/victoryi" directory within our project repository. By leveraging the ACOS quadruples generated from the Datast Preparation section, we have written starter code that employs the GPT API to convert these quadruples into summaries pertaining to a specific review aspect.

Although progress has been made, further work is required to enhance prompt engineering and identify the most suitable GPT model in order to optimize response time and improve the quality of the generated summaries. Currently, the produced summaries are largely functional, with minor instances of random textual artifacts, such as a randomly placed '#' within the aspect in the response. In the coming days, we will conduct additional prompt engineering to elicit more refined responses from the API.

1.3 Evaluation metrics

In order to assess the effectiveness of our model, we utilize two sets of metrics. The first set consists of widely-used metrics such as ROUGE and BERTScore for evaluating automated text summarization. The second consists of a novel set of metrics proposed by (Bhaskar et al., 2022) that evaluates the aspect-based review summaries generated by our model from three perspectives: faithfulness, factuality, and genericity. Specifically, failthfulness measures the degree to which the summary represents the consensus expressed in the original reviews; factuality assesses how strongly is the summary entailed by the reviews; genericity determines whether the summary is too generalized with certain words being overused. The faithfulness and factuality metrics are derived from the entailment score between the summary and original text using the entailment model from SummaC (Laban et al., 2021), whereas the genericity metric is derived by computing the Inverse Document Frequency (IDF) of the summaries.

However, there still remain challenges in implementing the new metrics. As the generated summaries often contain compound sentences with contrasting opinions, it is difficult for the entailment model to accurately infer the entailment score. Thus, we include a split-an-rephrase step by prompting the GPT model to split the summaries into sentences with simple propositions.

In the coming days, there are two tasks we aim to accomplish. First is to enhance prompt engineering for the split-and-rephrase operation. Second is to implement the other set of evaluation metrics, namely, ROUGE and BERTScore.

2 Risks and challenges

The implementation of our proposed review summarization system has encountered several challenges. Our goal is to develop a system that provides context to ratings, and we intend to evaluate its performance on new datasets obtained from Amazon or Yelp. One of the challenges we face is improving our system's summarization quality using GPT. We aim to enhance the quality of the summary responses generated by the GPT language model. We are currently working on improving our model to address this challenge. Another challenge is tuning the summary evaluations on these new datasets, given the unique characteristics of each dataset. A further challenge in our project is the difficulty of fine-tuning the T5 model for ACOS prediction. We intend to explore how well the T5 model can be adapted for this prediction task and evaluate its performance.

3 Plans to mitigate the risks and address the challenges

In order to mitigate the risks and address the challenges associated with the implementation of the proposed review summarization system, the project team proposes the following plan:
1. Enhancing the summarization quality: The team will allocate sufficient time to refine the prompt engineering and identify the most suitable GPT model to optimize response time and improve the quality of the generated summaries. The team will conduct additional prompt engineering to elicit more refined responses from the GPT API, including implementing a split-and-rephrase step to improve the entailment model's performance.
2. Evaluating system performance on new datasets: The team will evaluate the proposed review summarization system's performance on new datasets obtained from Amazon or Yelp. The team will analyze the unique characteristics of each dataset and modify the summarization parameters accordingly.
3. Tuning the summary evaluations: The team will implement the ROUGE and BERTScore evaluation metrics to assess the system's effectiveness in summarizing the review aspects. The team will also evaluate the performance of the proposed review summarization system using the novel metrics proposed by (Bhaskar et al., 2022), including faithfulness, factuality, and genericity.
4. Fine-tuning the T5 model: The team will explore the T5 model's adaptability for ACOS prediction and evaluate its performance. The team will allocate sufficient time to understand the T5 model's architecture and identify the best hyperparameters to improve the T5 model's performance.

To ensure effective project management, the team will establish clear communication channels between team members, provide regular project updates, and prioritize tasks based on their impact on project timelines and deliverables. Finally, the team will allocate sufficient time to test the proposed review summarization system's performance and identify any potential issues or risks. By implementing this plan, the project team can mitigate risks and address the challenges associated with the implementation of the proposed review summarization system, resulting in a successful outcome.

4 Individual Contributions

The members of the team have made significant contributions to the project. Pinyi Wang was responsible for the collection and analysis of the Laptop-ACOS and Restaurant-ACOS datasets. Victor Hui led the effort in implementing OpenAI's GPT API, while Xinyue Cui developed two sets of evaluation metrics for the system. 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.

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.