Self-Consistent Reasoning-based Aspect Sentiment Quad Prediction with Extract-Then-Assign Strategy

Jieyong Kim[1*] **Ryang Heo**[1*] **SeongKu Kang**[2] **Yongsik Seo**[1]
Jinyoung Yeo[1] **Dongha Lee**[1†]
Yonsei University[1] University of Illinois at Urbana Champaign[2]
{jieyong99,ryang1119,ndata,jinyeo,donalee}@yonsei.ac.kr
seongku@illinois.edu

Abstract

In the task of aspect sentiment quad prediction (ASQP), generative methods for predicting sentiment quads have shown promising results. However, they still suffer from imprecise predictions and limited interpretability, caused by data scarcity and inadequate modeling of the quadruplet composition process. In this paper, we propose Self-Consistent Reasoning-based Aspect sentiment quadruple Prediction (SCRAP), optimizing its model to generate reasonings and the corresponding sentiment quadruplets in sequence. SCRAP adopts the *Extract-Then-Assign* reasoning strategy, which closely mimics human cognition. In the end, SCRAP significantly improves the model's ability to handle complex reasoning tasks and correctly predict quadruplets through consistency voting, resulting in enhanced interpretability and accuracy in ASQP.[1]

1 Introduction

Aspect-based sentiment analysis (ABSA) refers to the task of identifying entity aspects and their associated sentiments (Pontiki et al., 2014). Among various ABSA tasks, the challenging task of predicting quadruplets, including aspect sentiment quad prediction (ASQP) (Zhang et al., 2021a) and aspect-category-opinion-sentiment (ACOS) (Cai et al., 2021), has garnered significant interest in that it can offer comprehensive aspect-level analysis. Specifically, a quadruplet consists of four sentiment elements: aspect term (*at*), opinion term (*ot*), aspect category (*ac*), and sentiment polarity (*sp*). Recent studies have developed powerful generative methods by fine-tuning language models (LMs) to sequentially generate sentiment quads (Zhang et al., 2021a; Hu et al., 2022; Gou et al., 2023).

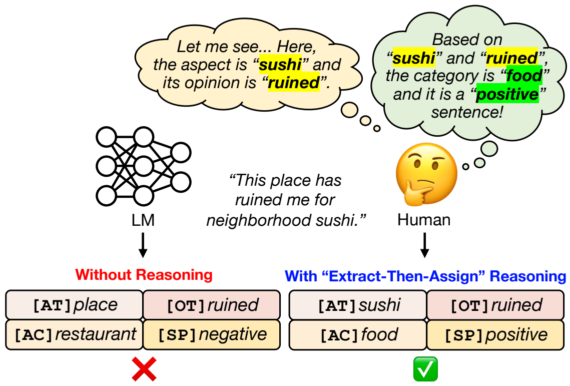

Figure 1: An illustrative example of Extract-Then-Assign reasoning process for ASQP task.

Although state-of-the-art generative methods for ASQP achieve promising accuracy, they are hindered by imprecise predictions and limited interpretability, stemming from data scarcity and inadequate modeling of the quadruplet composition process. In Figure 1 Left, the existing approach that directly decodes quads from a sentence, not only fails to make accurate predictions, but also struggles to discern the reasoning behind the quadruplet.

To tackle this challenge in ASQP, our approach introduces a two-step reasoning strategy, termed *Extract-Then-Assign*. This reasoning strategy initially extracts all aspect terms and opinion terms from an input sentence; subsequently, it assigns each aspect-opinion pair to both an aspect category and a sentiment polarity, utilizing predefined sets of categories and polarities. In Figure 1 Right, the process begins with a human extracting the information (*at*, *ot*) that is directly identifiable from the review sentence. Following this, the individual infers *ac* and *sp* based on the initially extracted components. We explicitly model these two steps of ASQP reasoning, which can account for the generation of complete quadruplets. This reasoning process, which closely mimics human cognition, allows us to enhance the accuracy and interpretability

[*] Equal contribution
[†] Corresponding author
[1] Codes and datasets are available at https://github.com/jieyong99/SCRAP

Findings of the Association for Computational Linguistics: ACL 2024, pages 7295–7303
August 11-16, 2024 ©2024 Association for Computational Linguistics

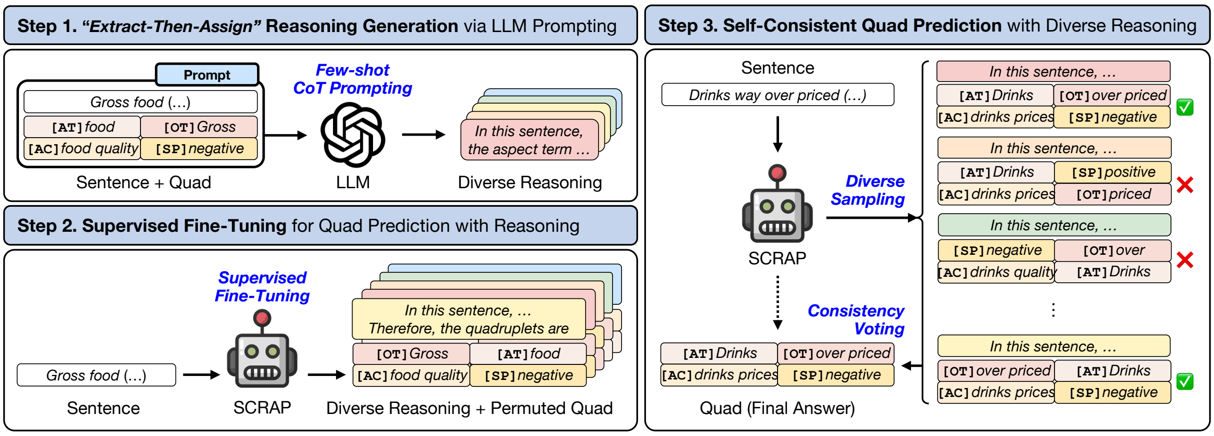

Figure 2: An overview of SCRAP which concurrently generates sentiment quads and the corresponding reasoning.

of quad prediction.

In this work, we propose a novel framework for **S**elf-**C**onsistent **R**easoning-based **A**spect-sentiment quad **P**rediction, named **SCRAP**, which predicts aspect sentiment quads based on the *Extract-Then-Assign* reasoning (Figure 2). The key idea is to distill the plausible reasoning ability from large language models (LLMs) into our ASQP model. To this end, we first collect diverse reasoning paths via Chain-of-Thought (CoT) prompting on LLMs (Wei et al., 2022), and then optimize our model to generate the reasoning followed by quad prediction. Furthermore, SCRAP aggregates its diverse reasoning outputs based on the self-consistency (Wang et al., 2022). This allows for filtering out noisy outputs and achieving more accurate quad predictions.

Extensive experiments on two ASQP benchmarks demonstrate that SCRAP significantly outperforms other state-of-the-art quad prediction models. Our reasoning process provides the explanation about the results and improves the prediction accuracy by understanding the inherent structure and relationships within the aspect-opinion pairs.

2 SCRAP Framework

2.1 Problem Formulation

Given an input sentence, aspect sentiment quad prediction (ASQP) aims to predict all aspect sentiment quads $\{(at, ot, ac, sp)\}$. The aspect term at and opinion term ot are detected within the sentence, while the aspect category ac and sentiment polarity sp are classified within their respective predefined sets. Please refer to Appendix A for the details.

2.2 ASQP Reasoning Generation

Extract-Then-Assign reasoning Mimicking the human cognition process, which first identifies terms and then infers their relations and semantics, we devise an *Extract-Then-Assign* reasoning strategy for ASQP. From an input sentence, it first extracts *at* and *ot* pairs, and subsequently infers the corresponding *ac* and *sp* by assigning them to elements within predefined sets of categories and polarities.

Reasoning generation with LLM We generate diverse reasoning paths using LLM based on the proposed reasoning strategy. We employ a few-shot Chain-of-Thought (CoT) prompting (Wei et al., 2022). Formally, given a sentence and its quadruplets, we generate N reasoning paths $\mathcal{R} = [r_1, r_2, ..., r_N]$ that explain how to reach the quadruplets from the sentence. Our prompt is designed to induce the *Extract-Then-Assign* process, facilitating the generation of plausible reasoning from the LLM. By leveraging the generated reasoning, we seek to provide rationales that guide and explain the ASQP task to our model, enhancing both accuracy and interpretability. The prompt can be found in the Appendix B.

2.3 Supervised Fine-Tuning

Target construction For each sentence, we construct the prediction targets by combining the generated reasoning with the ground-truth quadruplets, which serve as supervision for fine-tuning. This approach enables the model to learn reasoning ability from the LLM and grasp the intrinsic relationship between the reasoning and quadruplets.

For each input sequence, we construct multiple targets using various combinations of the reason-

Methods	Rest15			Rest16		
	Pre	**Rec**	**F1**	**Pre**	**Rec**	**F1**
TAS-BERT† (Wan et al., 2020)	44.24	28.66	34.78	48.65	39.68	43.71
Extract-Classify† (Cai et al., 2021)	35.64	37.25	36.42	38.40	50.93	43.77
GAS† (Zhang et al., 2021b)	45.31	46.70	45.98	54.54	57.62	56.04
Paraphrase† (Zhang et al., 2021a)	46.16	47.72	46.93	56.63	59.30	57.93
DLO† (Hu et al., 2022)	47.08	49.33	48.18	57.92	61.80	59.70
MvP† (Gou et al., 2023)	-	-	**51.04**	-	-	<u>60.39</u>
SCRAP (Ours)	55.45	45.41	<u>49.93</u>	69.59	56.70	**62.48**

Table 1: ASQP performance comparison. Backbone model: T5-Base. The best and second-best results are in **bold** and <u>underlined</u>, respectively. † indicates the results reported from their original papers.

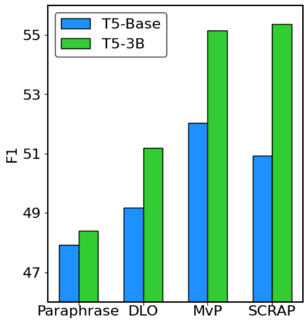

Figure 3: ASQP performance with T5-Base and T5-3B. Dataset: Rest15.

ing path r and the quadruplets q. For r, we use N different paths in $\mathcal{R}$. For q, we apply a data augmentation technique that uses P different permutations of elements in the quadruplet (Hu et al., 2022). Additional details and examples of target construction are provided in Appendix C.

Training Given an input sequence x, we train the model to predict the target y which consists of r and q. We fine-tune the sequence-to-sequence language model (Raffel et al., 2019) by minimizing the following negative log-likelihood loss, $\mathcal{L}_{NLL} = -\log p(y|x) = -\sum_{t=1}^{T} \log p(y_t|x, y_{<t})$, where T is the length of the target sequence y and $y_{<t}$ denotes previously generated tokens.

2.4 Self-Consistent Quad Prediction

At test time, the fine-tuned model proceeds with inference following the *Extract-Then-Assign* reasoning strategy, predicting the quadruplets along with the reasoning path. To mitigate the impact of noises on the reasoning process, we make the final prediction by consolidating multiple outputs based on self-consistency. Similar to (Wang et al., 2022) that samples diverse paths instead of only taking the greedy one, we sample K candidate outputs with diverse reasoning paths, and then identify the quadruplets consistently predicted by the model via consistency voting; this selects the quadruplets whose frequency exceeds a certain threshold.

3 Experiments

We experiment to answer the following questions:
RQ1: Does SCRAP outperform other baselines?
RQ2: How do diverse reasoning paths in SCRAP contribute to achieving higher accuracy?
RQ3: Does *Extract-Then-Assign* reasoning help to interpret quad prediction?

3.1 Experimental Settings

Dataset and evaluation metrics We use two datasets, i.e., Rest15 and Rest16, widely used for the ASQP task (Zhang et al., 2021a). For reasoning generation (Sec.2.3), we use ChatGPT (`gpt-3.5-turbo-16k`).[2] As the evaluation metric, we mainly employ the F1 score with precision (Pre) and recall (Rec). A predicted quad is considered as correct if and only if its all elements are exactly the same as the ground-truth ones.

Baselines We compare SCRAP with two discriminative methods, i.e., **TAS-BERT** (Wan et al., 2020) and **Extract-Classify** (Cai et al., 2021), as well as four competitive generative methods, i.e., **GAS** (Zhang et al., 2021b), **Paraphrase** (Zhang et al., 2021a), **DLO** (Hu et al., 2022), and **MvP** (Gou et al., 2023). For generative methods, we adopt T5-Base and T5-3B as backbone models. For SCRAP, we set $N = 16$ and $P = 5$ in common, $K = 20$ (T5-Base) and $K = 25$ (T5-3B) for Rest15, $K = 15$ for Rest16. Refer to Appendix D for implementation details.

3.2 Results and Discussion

SCRAP outperforms baseline methods (RQ1). Table 1 and Figure 3 present the ASQP performance of various methods. Overall, SCRAP achieves competitive performance, and outperforms the baseline methods for large backbone models. Specifically, in Figure 3, SCRAP shows higher effectiveness when applied to a larger model, exhibiting the largest performance gap between T5-Base and T5-3B. A larger model generally has higher reasoning capability (Li et al., 2023), making it more suitable for SCRAP which leverages

[2] `https://chat.openai.com/`

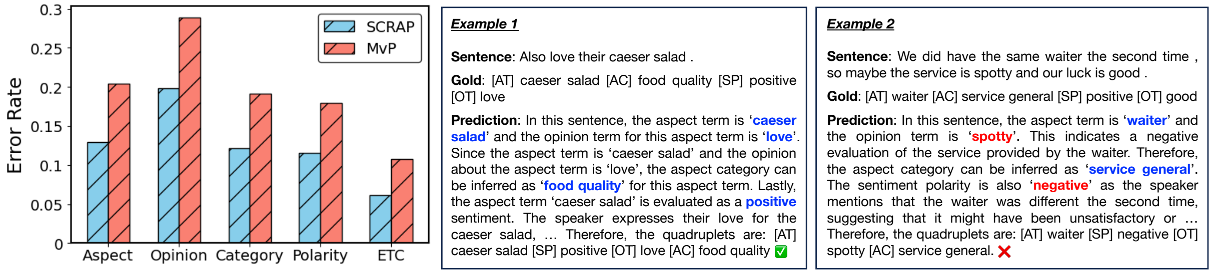

Figure 4: Error analysis and case study. **Left**: Analysis of prediction errors on the Rest16. We report the error rate for each element type of aspect sentiment quad. **Middle** and **Right**: The case study of SCRAP. We present the input sentence, gold quads, and the prediction made by SCRAP.

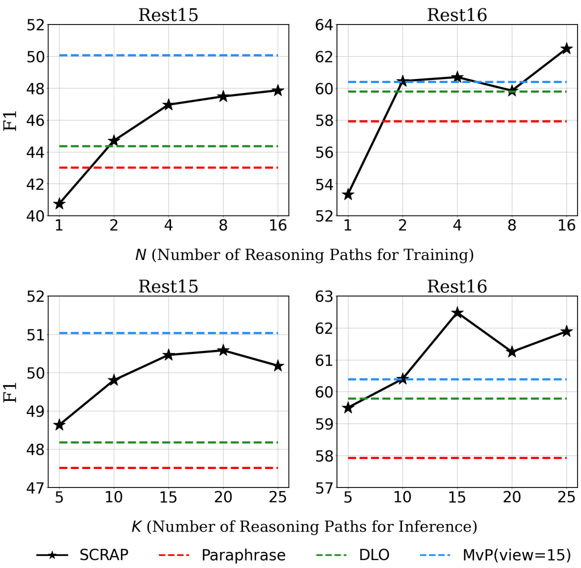

Figure 5: Effect of N and K, Backbone model: T5-Base.

reasoning for ASQP. Furthermore, SCRAP generally exhibits high precision, as it filters out inconsistent predictions through consistency voting. When the task is performed with the T5-Base on the Rest15 dataset, the performance is relatively low due to the small dataset size and the use of a model with relatively poor inference capabilities. Lastly, we analyze the prediction errors of SCRAP and the best competitor, MvP. The prediction error is calculated for each component as (*number of incorrect predictions*) / (*total number of predicted quads*). In Figure 4 Left, the overall error rate of SCRAP is notably lower than MvP.

Impact of the number of reasoning paths (RQ2). We investigate the performance of SCRAP with varying N and K, which control the number of reasoning paths used for the training and inference, respectively. In Figure 5, we observe that SCRAP generally achieves higher performance with a greater number of reasoning paths with respect to both N and K, and the best performance is achieved by leveraging multiple paths for both training and inference. These results show that diverse reasoning for prediction is indeed beneficial in more accurate quad predictions. Nevertheless, if K grows excessively, a considerable quantity of incomplete inferences are produced, leading to adverse effects on quad predictions and consequent performance deterioration. Therefore, it is crucial to determine the optimal value of multiple paths K for inference.

***Extract-Then-Assign* reasoning offers interpretability for quad prediction (RQ3).** Figure 4 Mid and Right provide the case study. In *Example 1*, the model predicts the correct quad, and it is possible to interpret the process through which the correct answer was reached. Moreover, even if the model fails to predict the correct quad, we can still understand how the prediction failed based on the *Extract-Then-Assign* reasoning process. In *Example 2*, the model fails to extract the correct opinion term, and this leads to inaccurate polarity prediction. This interpretability helps to better understand the model behavior, which is an important strength of SCRAP. It demonstrates that the *Extract-Then-Assign* reasoning strategy is simple yet effective within the SCRAP framework.

4 Related Work

Aspect Sentiment Quad Prediction Many existing studies have focused on discriminative methods. Early works tried jointly detecting target-aspect-sentiment (Wan et al., 2020) or conducting ACOS with two-stage pipelines (Cai et al., 2020). Beginning with Zhang et al. (2021b), recent studies have started using generative methods. Zhang et al. (2021a) transforms outputs into natural language sentences, Hu et al. (2022) introduce data augmentation that uses various permutations of quad elements, and Gou et al. (2023) align training and inference with multi-view prompting.

Chain-of-Thought (CoT) Distillation CoT prompting has shown high effectiveness in inducing models to generate reasoning before reaching an answer (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022). Recent work has focused on distilling the reasoning capabilities of LLMs to smaller LMs (Ho et al., 2022; Magister et al., 2022; Hsieh et al., 2023); they elicit rationales for the predictions from LLMs and utilize them to train the LMs, effectively improving their performance.

5 Conclusion

This paper aims to enhance quad prediction in ASQP using *Extract-Then-Assign*, which is the two-step reasoning strategy. To this end, we propose a SCRAP framework, which generates diverse predictions utilizing the *Extract-Then-Assign* and selects the final answer by filtering the inconsistent answers through consistency voting. Our framework is the first method to integrate reasoning into the ABSA task, not only achieving state-of-the-art performance, but also significantly enhancing the interpretability of the outputs produced by the model. This confirms its efficacy in predicting quadruplets through reasoning.

6 Limitations

Despite achieving state-of-the-art performance, our study has three limitations. Firstly, our current reasoning structure, designed to mimic human cognition, may not be the most advanced or optimal for the ASQP task. There could exist a more sophisticated reasoning structure that further enhances the task performance. Secondly, the effectiveness of our approach is affected by the size of the model to some extent. With small models having limited reasoning capabilities, our method may not exhibit satisfactory performance. Lastly, our study incurs higher computational costs for training and inference compared to previous studies, as it additionally leverages reasoning.

7 Ethical Statment

We utilize datasets that are widely recognized and previously employed in the scientific community, maintaining transparency and integrity in our experiments. Our methodologies and findings do not inflict harm upon any individuals or groups. We are cognizant of the potential biases in sentiment polarity predictions arising from the use of large pre-trained language models, as these models may mirror existing societal biases found in their training corpora (Tan and Celis, 2019). We acknowledge the importance of ongoing efforts to mitigate such biases. Furthermore, we underscore the necessity of continuous monitoring and evaluation to ensure that our smaller downstream models do not replicate or amplify the biases inherent in their larger language model counterparts.

Acknowledgements

This work was supported by the IITP grant funded by the Korea government (MSIT) (No. RS-2020-II201361) and the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00244689).

References

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *International Conference on Computational Linguistics*.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Annual Meeting of the Association for Computational Linguistics*.

Zhibin Gou, Qi Guo, and Yujiu Yang. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. In *Annual Meeting of the Association for Computational Linguistics*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. In *Annual Meeting of the Association for Computational Linguistics*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander J. Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *ArXiv*, abs/2305.02301.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Mengting Hu, Yike Wu, H. Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Conference on Empirical Methods in Natural Language Processing*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. *ArXiv*, abs/2306.14050.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *ArXiv*, abs/2212.08410.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Yi Chern Tan and Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *ArXiv*, abs/1911.01485.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *AAAI Conference on Artificial Intelligence*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Conference on Empirical Methods in Natural Language Processing*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics*.

A Problem Formulation Details

The aspect term includes the possibility of being null, which denotes cases where it is not explicitly mentioned, and this is represented by "*NULL*". In this work, the aspect category *ac* is classified as an element within the predefined set: {"food prices", "food style_options", "service general", "drinks prices", "ambience general", "drinks quality", "location general", "restaurant prices", "restaurant general", "drinks style_options", "food general", "restaurant miscellaneous", "food quality"}. The sentiment polarity *sp* is categorized into one of the three sentiment classes: {"*positive*", "*neutral*" or "*negative*"}, each signifying the respective emotional disposition conveyed.

B Reasoning Generation Details

We use ChatGPT (`gpt-3.5-turbo-16k`) to generate reasoning paths for each training sample, and use some or all of them for fine-tuning purposes. We carefully design the prompt to induce the *Extract-Then-Assign* process, facilitating the generation of plausible rationales for ASQP. We present examples of our prompts in Table 3.

C Target Construction Details

We construct the prediction targets by combining the generated reasoning with the ground-truth quadruplets. To represent the quadruplets, we use special markers: `[AT]`, `[OT]`, `[AC]`, `[SP]`, which respectively denote *at*, *ot*, *ac*, *sp*, as done in (Hu et al., 2022). We also employ the data augmentation technique that uses various permutations of elements in the quadruplet (Hu et al., 2022). The element permutations are ranked based on the prediction entropy of pre-trained T5-Base, and we use P different permutations with minimal entropy for each quadruplet. If there are multiple quadruplets for a single sentence, we concatenate them using a special symbol `[SSEP]` (2021a).

We combine the reasoning and the quadruplets with a connecting expression *'Therefore, the quadruplets are:'*. An example of an input sentence and the constructed target is provided in Table 2.

D Experiment Details

D.1 Software and Hardware

We use Pytorch to implement all the models (Python 3.8). Our experiments are conducted with a single NVIDIA A6000 GPU with 48GB of RAM.

D.2 Implementation Details

For training, the batch size is set to 16 for T5-Base and 2 for T5-3B. The number of training epochs is set to 20 and the learning rate is set to $3e-5$ for all experiments. To accelerate training and reduce GPU memory consumption, we apply LoRA (Hu et al., 2021) for T5-3B. For the self-consistent quad prediction, we independently generate up to 25 predictions (K) from the decoder, utilizing the temperature sampling (Wang et al., 2022). We explore temperatures ranging from 0.7 to 1.5. We observed that when the temperature is too low (e.g. $T \leq 0.7$), it tends to produce identical or similar reasoning for the input sentence. Conversely, when the temperature is too high (e.g., $T \geq 1.5$), the generated reasoning is often incomplete, adversely affecting quad prediction. We empirically choose the appropriate temperature as T = 1.3 in our experiments.

Input-Target example
Input: we have gone for dinner only a few times but the same great quality and service is given .   **Target:** In this sentence, the speaker mentions their experience of going for dinner a few times. They highlight the consistent great quality and service they have received. For the first quadruplet, the aspect term is ’service’ and the opinion term is ’great’. This indicates a positive evaluation of the service provided by the restaurant. Therefore, the aspect category can be inferred as ’service general’ and the sentiment polarity is positive. For the second quadruplet, the aspect term is ’dinner’ and the opinion term is ’great quality’. This implies a positive assessment of the quality of the food during dinner. Hence, the aspect category can be inferred as ’food quality’ and the sentiment polarity is positive. The speaker’s statement overall conveys their positive experience, emphasizing the consistent great quality and service they have received during their visits for dinner. Therefore, the quadruplets are: [SP] positive [AT] dinner [OT] great quality [AC] food quality [SSEP] [SP] positive [AT] service [OT] great [AC] service general

Table 2: Input-Target example for ASQP

***Extract-Then-Assign* Prompt**
[Task Description] I am performing the ASQP task, which is the Subtask of ABSA. From now on, if I give you a sentence and a quadruplet, create a Reasoning for it. When creating, create to satisfy all of the following conditions: When proceeding with inference, extract the aspect term and the option term first and infer the aspect category and sentimental polarity based on them. When extracting aspect term and option term, an aspect category is judged by a combination of aspect term and opinion term, and sentimental polarity is judged by comprehensively considering everything. And don't mention each element first, explain the reason first, and then create a rationale that mentions the element. At this time, do not number each element, but configure it to naturally lead to one paragraph. But if there are more than two quadruplets, please organize the description for each quadruplet. And please create a detailed description of each element in the composition of rationale. From now on, I'll give you sentences and quadruplet sets as input. Here are possible aspect category set: ['food prices', 'food style_options', 'service general', 'drinks prices', 'ambience general', 'drinks quality', 'location general', 'restaurant prices', 'restaurant general', 'drinks style_options', 'food general', 'restaurant miscellaneous', 'food quality']. Here are possible sentiment polarity set: ['positive','negative','neutral']. **[Example 1]** **Text:** The fried dumplings are GREAT !####[['fried dumplings', 'food quality', 'positive', 'GREAT']] **Reasoning:** In this sentence, the aspect term is 'fried dumplings' and the opinion term for this aspect term is 'GREAT'. Since the aspect term is 'fried dumplings' and the opinion about the aspect term is 'GREAT', the aspect category can be inferred as 'food quality' for this aspect term. Lastly, the aspect term 'fried dumplings' is evaluated as a opinion of 'GREAT'. When it comes to food, the opinion 'GREAT' suggests that the food is delicious, which is evaluated as a positive sentiment. **[Example 2]** **Text:** It's one of our favorite places to eat in NY.####[['NULL', 'restaurant general', 'positive', 'favorite']] **Reasoning:** In this sentence, there is no specific aspect term mentioned explicitly. So the aspect term is 'NULL' and the opinion term for this aspect term is 'favorite'. The aspect category could be inferred as 'restaurant general' as the speaker is expressing a general sentiment about the restaurant rather than a specific feature or component. Lastly, by referring to the restaurant as a 'favorite', the speaker implies a positive sentiment polarity. **[Example 3]** **Text:** It is very overpriced and not very tasty .####[['NULL', 'food quality', 'negative', 'not very tasty'], ['NULL', 'food prices', 'negative', 'overpriced']] **Reasoning:** In the sentence, there are two different evaluations made but none of them explicitly mentions the specific aspect term. So, for both evaluations, the aspect term is 'NULL'. For the first quadruplet, the opinion term is 'not very tasty'. This is a negative evaluation of the food quality, hence 'food quality' can be inferred as the aspect category and the sentiment polarity is negative. In the second quadruplet, the opinion term is 'overpriced'. This term is often used to describe something that is too expensive or not worth the price. Therefore, the aspect category can be inferred as 'food prices' and since the speaker is expressing a negative sentiment about the price, the sentiment polarity is negative. **[Example 4]** **Text:** The service was friendly and the atmosphere was casual .####[['service', 'service general', 'positive', 'friendly'], ['atmosphere', 'ambience general', 'neutral', 'casual']] **Reasoning:** The sentence discusses two aspects - 'service' and 'atmosphere'. For the first quadruplet, the aspect term is 'service' and the opinion term is 'friendly'. This is a positive evaluation of the service provided by the restaurant, hence 'service general' can be inferred as the aspect category and the sentiment polarity is positive. In the second quadruplet, the aspect term is 'atmosphere' and the opinion term is 'casual'. This term is used to describe the general ambience of the restaurant. Therefore, the aspect category can be inferred as 'ambience general'. As the term 'casual' is neutral and doesn't indicate any positive or negative sentiment, the sentiment polarity is neutral. **[Example 5]** **Text:** Rude service , medicore food ... there are tons of restaurants in NY ... stay away from this one####[['service', 'service general', 'negative', 'Rude'], ['food', 'food quality', 'neutral', 'medicore'], ['NULL', 'restaurant general', 'negative', 'stay away']] **Reasoning:** The sentence discusses three aspects - 'service', 'food', and the general experience at the restaurant (NULL). For the first quadruplet, the aspect term is 'service' and the opinion term is 'Rude'. This is a negative assessment of the service provided by the restaurant, hence 'service general' can be inferred as the aspect category and the sentiment polarity is negative. In the second quadruplet, the aspect term is 'food' and the opinion term is 'medicore'. This term is used to describe the quality of the food at the restaurant. Therefore, the aspect category can be inferred as 'food quality'. As the term 'medicore' is neutral and doesn't indicate any positive or negative sentiment, the sentiment polarity is neutral. In the third quadruplet, there is no specific aspect term mentioned, so the aspect term is 'NULL'. The opinion term is 'stay away'. This is a negative sentiment about the restaurant in general, hence 'restaurant general' can be inferred as the aspect category and the sentiment polarity is negative.

Table 3: The prompt for *Extract-Then-Assign* on ASQP.