Exploring Rollback Inference for Aspect-based Sentiment Analysis

Anonymous ACL submission

Abstract

With the giant help from pre-trained large language models (LLMs), templated sequence of how to organize the aspect-level elements become the hottest research target while only a few of them move their steps to inference, not to mention utilizing the semantic connection between aspect-level elements during it. We argue that, compared with the high computational cost methods of training language models, considering the inference process can also bring us potential benefits. Motivated by this, we propose *rollback inference* for aspect-based sentiment analysis, which can boost the performance of fine-tuned LLMs with a tiny cost, and adapt to various language models. Specifically, We first propose a novel entropy-based rollback inference framework that manipulates multi-reasoning and voting over the uncertain parts of the sequence with model's self-consistency. We then explore the possibility of capturing the correlations among elements during inference with a set of rollback strategies. Extensive experiments in several benchmarks underscore the robustness and effectiveness of our proposed rollback strategies and the value of the semantic connections in inference.

1 Introduction

Aspect-based sentiment analysis (ABSA) has garnered growing interest in the community, encompassing four subtasks: aspect term extraction, opinion term extraction, aspect term category classification, and aspect-level sentiment classification. The initial two subtasks focus on extracting the aspect term and the opinion term present in the sentence. The objectives of the last two subtasks are to identify the category and sentiment polarity related to the extracted aspect term.

The sentiment quadruple extraction task, which is composed of four subtasks, poses a significant challenge for traditional classification-based models due to its complexity. In response to this challenge, recent studies have adopted a unified generative approach that circumvents the need for explicit modelling of the ABSA problem. This approach treats either the class index (Yan et al., 2021), or the desired sentiment element sequence (Zhang et al., 2021b,a; Bao et al., 2022), as the target output of the generative model. By doing so, these studies aim to simplify the overall task and improve its effectiveness.

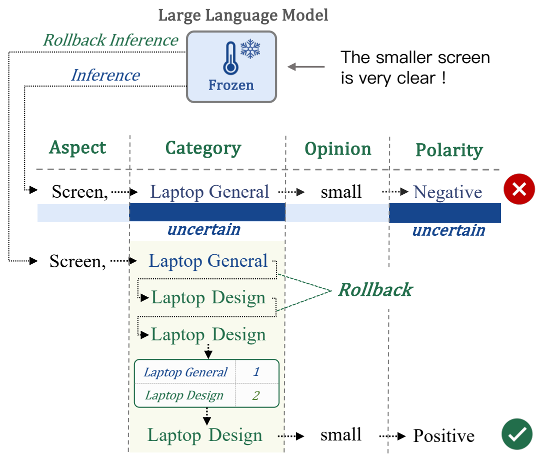

Figure 1: Example of proposed rollback inference framework.

However, most previous studies have concentrated on enhancing the training phase of generative models for sentiment analysis (Zhang et al., 2021b; Bao et al., 2022; Hu et al., 2022), simply adopting greedy search from left to right and neglecting the significance of the inference stage. As a result, the majority of these models rely on post-processing steps to ensure structural integrity (Bao et al., 2022, 2023b). In addition, these models fail to grasp the correlations among sentiment elements during inference (Hu et al., 2022), thus compromising the comprehensiveness of the sentiment analysis.

In this study, we direct our attention to the inference stage of sentiment generation models. We observe that, once the model is uncertain about one element, it tends to perform a similar attitude on the other elements in the same quadruple that are semantically connected. As shown in Figure 1, uncertainty is reported for both the category of aspect and polarity.

Motivated by this, we introduce a novel self-consistency rollback inference framework along with a set of rollback strategies to better capture the correlations among sentiment elements during inference and improve its overall effectiveness. As illustrated in Figure 1, we employ an entropy-based mechanism to assess the uncertainty of sentiment elements during inference. When an element is deemed uncertain based on its entropy score, we launch a rollback procedure. This rollback is performed on a specific span determined by our proposed rollback strategies, resampling the span multiple times to get diverse results. Finally, we employ a majority vote mechanism to determine the final results for the rollback span.

The detailed evaluation shows that our model significantly advances the state-of-the-art performance on several benchmark datasets. In addition, the empirical studies also indicate that the proposed rollback inference strategy is more effective than other inference strategies.

2 Related Work

Generative ABSA: Research on ABSA typically follows a progression from addressing individual sub-tasks to dealing with their intricate combinations. The initial focus is often on predicting a single sentiment element (Wang et al., 2021; Hu et al., 2019; Tang et al., 2016; Chen et al., 2022; Liu et al., 2021; Seoh et al., 2021; Zhang et al., 2022). Many studies also delve into exploring the joint extraction of sentiment elements (Xu et al., 2020; Li et al., 2022; Bao et al., 2023a; Zhang and Qian, 2020).

More recently, there are some attempts to tackle ABSA problem in a generative manner (Zhang et al., 2021a), either treating the class index (Yan et al., 2021) or the desired sentiment element sequence (Zhang et al., 2021b) as the target of the generation model. For example, Yan et al. (2021) employed a sequence-to-sequence pre-trained model to generate the sequence of aspect terms and opinion words directly. Meanwhile, Zhang et al. (2021a) proposed a paraphrasing model that utilized the knowledge of the pre-trained model via casting the original task to a paraphrase generation process. In addition, Bao et al. (2022) addressed the importance of correlations among sentiment elements, and proposed an opinion tree generation model to jointly detect all sentiment elements in a tree structure.

Decoding Strategies for LLMs: Multiple decoding strategies for language models have been proposed on general tasks to explicitly promote diversity in the decoding process in the literature, e.g., temperature sampling (Ackley et al., 1985; Ficler and Goldberg, 2017), top-k sampling (Radford et al., 2019; Holtzman et al., 2018; Fan et al., 2018), nucleus sampling (Holtzman et al., 2020). Besides, for improving accuracy, Self-consistency(COT-SC) decoding (Wang et al., 2023) has been proposed to explore multiple different ways of thinking leading to its unique correct answer.

However, the regeneration of the entire sequence in COT-SC is not applicable to the ABSA task as the reasoning associations between quadruples are not as strong as the general reasoning process. Huang et al. (2023) controlled text generation with arbitrary plugins during inference, which however requires to be trained separately. Gou et al. (2023) employed a majority vote decoding over different template orders, treating elements equally without semantic distinction. Hu et al. (2023) somehow proposed marginalized unlikelihood learning to suppress the uncertainty-aware mistake tokens.

Unlike previous works that often require complex pre-processing or post-processing steps, our method does not need such procedures. Instead, it easily integrates with fine-tuned language models, achieving significant improvements with only a minor increase in inference time. This makes our strategy a practical and efficient solution for enhancing sentiment analysis during the inference stage.

3 Aspect-based Sentiment Analysis with Rollback Inference

As shown in Figure 2, we introduce a novel *rollback inference framework* for generative aspect-based sentiment analysis.

To begin, we first fine-tune a large language model and freeze its parameters before entering the inference stage. Next, during inference, we propose an entropy-based mechanism to assess the uncertainty of sentiment elements and resample the uncertain span (detailed in Section 4) multiple times to get diverse results to construct the candidates

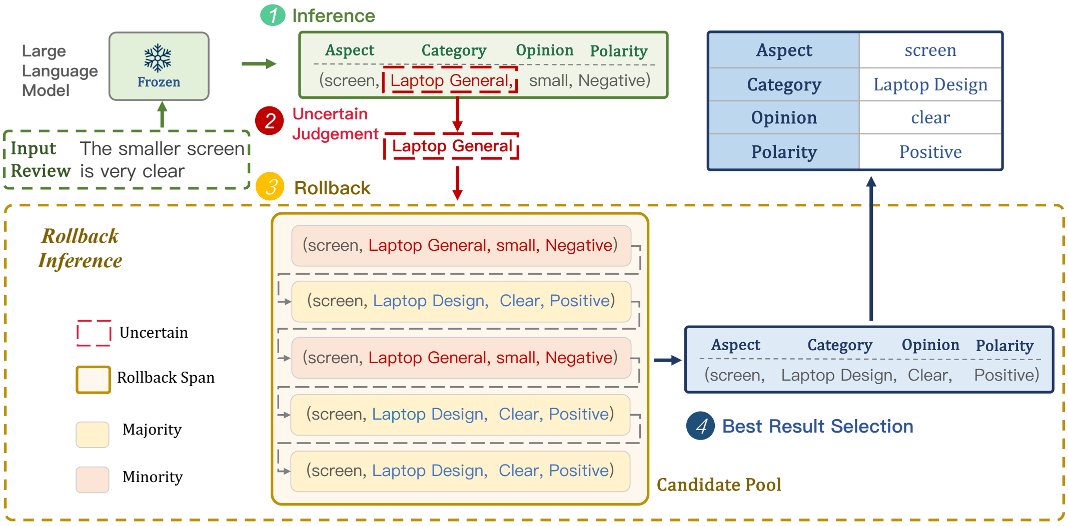

Figure 2: Overview of proposed rollback inference framework.

pool. Finally, we obtain a final self-consistency result for the rollback span with a majority vote mechanism over the candidates.

3.1 Generative Aspect-based Sentiment Analysis

In this study, we fine-tune the pre-trained large language model LLaMA (Touvron et al., 2023) as our foundation. This model receives a review sentence as input and produces sentiment quadruples as output.

Given the token sequence $x = x_1, ..., x_{|x|}$ as input, the model outputs the linearized representation $y = y_1, ..., y_{|y|}$. The decoder predicts the output sequence token-by-token. At the i-th step of generation, the decoder predicts the i-th token y_i in the linearized form, and decoder state h_i^d as:

$$y_i, h_i^d = ([h_1^d, ..., h_{i-1}^d], y_{i-1}) \quad (1)$$

The conditional probability of the whole output sequence $p(y|x)$ is progressively combined by the probability of each step $p(y_i|y_{<i}, x)$:

$$p(y|x) = \prod_{i=1}^{|y|} p(y_i|y_{<i}, x) \quad (2)$$

where $y_{<i} = y_1 ... y_{i-1}$, and $p(y_i|y_{<i}, x)$ are the probabilities over target vocabulary V.

The objective functions is to maximize the output target sequence X_T probability given the review sentence X_O. Therefore, we optimize the negative log-likelihood loss function:

$$\mathcal{L} = -\frac{1}{|\tau|} \sum_{(X_O, X_T) \in \tau} \log p(X_T|X_O; \theta) \quad (3)$$

where θ is the model parameters, and (X_O, X_T) is a (*sentence, tree*) pair in training set τ, then

$$\log p(X_T|X_O; \theta) = \\ = \sum_{i=1}^{n} \log p(x_T^i | x_T^1, x_T^2, ... x_T^{i-1}, X_O; \theta) \quad (4)$$

where $p(x_T^i | x_T^1, x_T^2, ... x_T^{i-1}, X_O; \theta)$ is calculated by the decoder.

3.2 Uncertain Element Judgement

During the inference stage of our generative aspect-based sentiment analysis model, we introduce an uncertain judgement mechanism to address elements that need rollback. This mechanism is triggered whenever the model generates a token with low confidence. Instead of accepting this uncertain token, we rollback to a previous state and regenerate the semantically connected span.

To quantify the model's certainty, we adopt information entropy as a metric. Specifically, for each generation step i, we calculate the entropy E_i using the formula:

$$E_i = -\sum_{j}^{M} P(x_j) log(P(x_j)) \quad (5)$$

Here, $P(x_j)$ represents the output probability of the j-th token in the vocabulary, and M denotes the vocabulary size. A higher entropy E_i indicates that the model is less certain about its choice at step i.

When the entropy exceeds a predefined threshold, we consider the model to be uncertain and

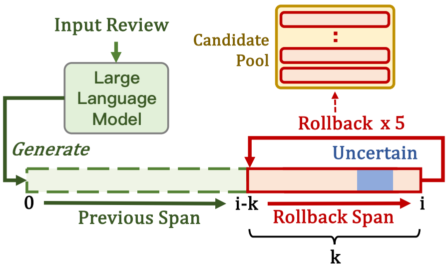

Figure 3: Example of rollback inference.

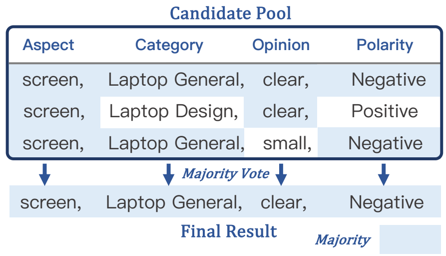

Figure 4: Illustration of the best result selection.

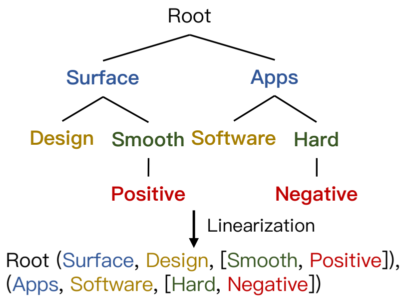

Figure 5: Example of the opinion tree structure.

initiate the rollback process. This involves revisiting the semantically connected span and potentially generating a new set of candidates. The most confident candidate is then selected as the new output, ensuring that the model's predictions are both self-consistent and reliable.

3.3 Rollback Inference

When an element is judged to be uncertain during the generation process, we employ a rollback strategy to revisit the corresponding span related to that element as shown in Figure 3. We adopt sampling in rollback inference, which choosing next token randomly with probability distribution instead of greedy search to ensure the diversity of candidates. Since the span of rollback is the key issue of this stage, we will discuss it in the next section.

We first generate sequence normally if there are no elements judged uncertain (green bar in Figure 3). Once an element is judged uncertain as the blue printed, we would like to rollback the corresponding span (printed red) related to it. Assuming we rollback at step i with a length k (determined by specific strategy), we would retreat the steps back to step $i - k$ and resample the following sequence to step i multiple times, the rollbacked span would be served as a candidate.

By rolling back multiple times, we can construct a pool of candidates for the uncertain sub-sequence. This pool provides the model with multiple options to choose from, increasing the chances of finding a more accurate and self-consistent prediction. The final prediction is then selected from this pool based on a predefined criterion, such as the highest confidence score or majority voting.

3.4 Best Result Selection

After constructing a pool of candidates for the uncertain sub-sequence, we proceed to select the best result from among these candidates as the final output. As illustrated in Figure 4, our approach involves dividing each candidate into its constituent sentiment elements. We then tally the votes for each element by counting the number of occurrences of its type (e.g., aspect, opinion, and polarity).

The sentiment element with the highest number of votes is subsequently selected as the final result. This majority voting mechanism allows us to leverage the collective wisdom of the model's predictions, thereby increasing its confidence in the chosen output, especially for uncertain sub-sequences.

4 Rollback Inference Strategies

In this section, we first introduce the utilization of an opinion tree structure as a means to systematically organize and represent various sentiment elements. This tree structure serves as the backbone for rollback inference strategies. We then introduce different rollback inference strategies designed to select suitable candidates for uncertain sub-sequences in it.

4.1 Opinion Tree Construction

As shown in Figure 5, the opinion tree is hierarchically structured, beginning with a root node. The children of this root node are quadruple sub-trees, each rooted at an aspect node. These aspect nodes are then connected to category and opinion nodes, which together form the branches of the sub-tree.

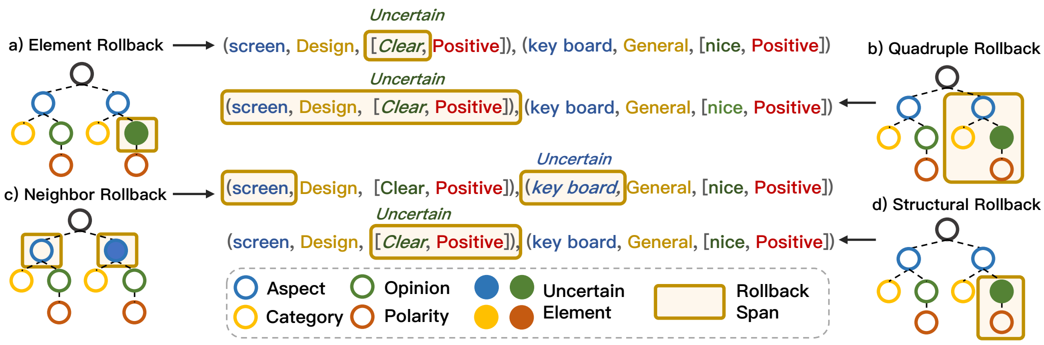

Figure 6: Illustration of the specific rollback span of proposed strategies.

Polarity nodes are positioned as the successors of the corresponding opinion nodes, completing the structural representation of sentiment elements.

The linearization of this tree structure results in the final target sequence, which preserves the hierarchical relationships and semantic connections among sentiment elements.

4.2 Element Rollback

Element Rollback inference (ER) represents a fundamental rollback strategy characterized by its narrow rollback span, which minimizes the additional inference time required.

As illustrated in Figure 6(a), when a token within an element is determined to be uncertain, the element would be regarded as the rollback span and underwent rollback multiple times to construct a pool of candidates. As it rollbacks each single token, Element Rollback can be applied on any generative tasks, irrelevant to the form.

4.3 Quadruple Rollback

Quadruple Rollback inference (QR) is an intuitive strategy that recognizes the natural co-relation among the elements within a quadruple. This approach designs a holistic packaging strategy to address the entire quadruple as a unified entity.

As shown in Figure 6(b), when a token within the sub-sequence of a quadruple is deemed uncertain, the entire quadruple undergoes rollback. This means that instead of focusing solely on the uncertain token, Quadruple Rollback considers the broader context provided by the other elements within the quadruple.

4.4 Neighbor Rollback

Neighbor Rollback inference (NR) is a strategy tailored to the structural formation of data, operating under the assumption that the neighbors (or sibling nodes) of an uncertain element may be influenced by its uncertainty.

As illustrated in Figure 6(c), when a token within an element of a quadruple is determined to be uncertain, Neighbor Rollback targets the siblings of this element as the rollback span. This means that instead of rolling back the entire quadruple or just the single uncertain element, Neighbor Rollback focuses on the immediate vicinity of the uncertain element.

4.5 Structural Rollback

In the context of structural opinion trees, the parent node (also known as the root node of a sub-tree) serves as the semantic foundation for the child nodes that originate from it. The uncertainty associated with a parent node has the potential to propagate throughout the entire sub-tree rooted at that node due to the shared semantic connections.

Recognizing this, we have developed a Structural Rollback inference strategy (SR) tailored to the inherent properties of the opinion tree. This strategy aims to address uncertainty at its source, the parent node, and mitigate its impact on the broader sub-tree structure.

As shown in Figure 6(d), during the inference process, if a token within a sentiment node of the opinion tree is deemed uncertain, the inference continues uninterrupted until it reaches the terminus of the sub-tree rooted at that sentiment node. Once this point is reached, the entire sub-tree undergoes multiple rollbacks initiated by the framework.

5 Experiments

In this section, we introduce the datasets used for evaluation and the baseline methods employed for comparison. We then report the experimental results and analyze the effectiveness of the proposed model with different factors.

Method	Restaurant			Laptop			Phone		
	P	R	F1	P	R	F1	P	R	F1
JET	0.5981	0.2894	0.3901	0.4452	0.1625	0.2381	0.3845	0.2213	0.2809
TAS-BERT	0.2629	0.4629	0.3353	0.4715	0.1922	0.2731	0.3453	0.2207	0.2693
Extract-Classify	0.3854	0.5296	0.4461	0.4556	0.2948	0.3580	0.3128	0.3323	0.3223
Seq2Path	0.6029	0.5961	0.5995	0.4448	0.4375	0.4411	0.5263	0.4994	0.5125
OTG	0.6191	0.6085	0.6164	0.4395	0.4383	0.4394	0.5302	0.5659	0.5474
One-ASQP	**0.6591**	0.5624	0.6069	0.4380	0.3954	0.4156	**0.5742**	0.5096	0.5400
GAS	0.6069	0.5852	0.5959	0.4160	0.4275	0.4217	0.5072	0.4815	0.4940
Paraphrase	0.5898	0.5911	0.5904	0.4177	**0.4504**	0.4334	0.4672	0.4984	0.4832
DLO	0.5904	0.6029	0.5966	0.4359	0.4367	0.4363	0.5451	0.5173	0.5308
ChatGPT	0.5014	0.3625	0.4207	**0.4492**	0.3123	0.3541	0.4514	0.4627	0.4569
LLaMA	0.6213	0.6024	0.6117	0.4334	0.4201	0.4266	0.5314	0.5478	0.5394
Ours	0.6585	**0.6197**	**0.6382**	0.4470	0.4417	**0.4443**	0.5387	**0.5709**	**0.5543**

Table 1: Results in ACOS and en-Phone, we report the performance of our proposed model with structure rollback.

Method	Rest15			Rest16		
	P	R	F1	P	R	F1
HGCN-BERT+BERT-TFM*	0.2555	0.2201	0.2365	0.2740	0.2641	0.2690
TASO-BERT-CRF*	0.4424	0.2866	0.3478	0.4865	0.3968	0.4371
GAS*	0.4531	0.4670	0.4598	0.5454	0.5762	0.5604
Paraphrase*	0.4616	0.4772	0.4693	0.5663	0.5930	0.5793
DLO*	0.4708	0.4933	0.4818	0.5792	**0.6180**	0.5979
Ours	**0.4968**	**0.5082**	**0.5024**	**0.5937**	0.6153	**0.6043**

Table 2: Results in Rest15/16, we report the performance of our proposed model with structure rollback. The results of baseline methods, marked with *, are obtained from this work (Hu et al., 2022)

5.1 Dataset and Experiment Setting

In this study, we use restaurant and laptop domains in ACOS dataset (Cai et al., 2021) and phone domain in Zhou et al. (2023)'s dataset for our experiments. We also include Rest15 and Rest16 datasets(Zhang et al., 2021a) for a comprehensive comparison.

For our opinion tree generation model, we employ LLaMA-2-7B[1] and LoRA fine-tune the adapter parameters. We tune the parameters of our models by grid searching on the validation dataset. We fine-tune the model with 20 epochs and save the model parameters for inference. The LoRA alpha is set to 128 and LoRA rank is set to 64.

The model parameters are optimized by Adam (Kingma and Ba, 2015), the learning rate of fine-tuning is 5e-5. The batch size is set to 4 with a cut-off length of 1024. The LoRA adapter would be merged with the original LLaMA-2-7B parameters and freeze during the inference process. During inference, we do sampling and set the entropy threshold to 0.6, rollback times to 5, top K to 2, temperature to 0.95 with beam size 1 and average the 5 runs as the final result. Our experiments are carried out with an Nvidia RTX4090.

[1]LLaMA-2-7B-Chat,https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

In evaluation, a quadruple is viewed as correct if and only if the four elements, as well as their combination, are exactly the same as those in the gold quadruple. On this basis, we calculate the Precision and Recall, and use F1 score as the final evaluation metric for aspect sentiment quadruple extraction (Cai et al., 2021; Zhang et al., 2021a).

5.2 Main Results

In Table 1 and 2, we present a comprehensive comparison of our proposed model with various state-of-the-art baselines. These baselines include both extraction-based methods and generative models, as well as large language models.

Extraction-based methods, such as JET (Xu et al., 2020), TAS-BERT (Wan et al., 2020; Zhang et al., 2021a),HGCN-BERT+BERT (Zhang et al., 2021a), and Extract-Classify (Cai et al., 2021), typically rely on identifying relevant spans within the input text to extract sentiment quadruples. On the other hand, generative models, such as GAS (Zhang et al., 2021b), Paraphrase (Zhang et al., 2021a), DLO (Hu et al., 2022), Seq2Path (Mao et al., 2022), OTG (Bao et al., 2022)[2], and One-ASQP (Zhou et al., 2023), aim to

[2]We adopt the OTG performance without external resource pre-training for fair comparison.

Method	Manner	Time(s)	Restaurant	Laptop	en-Phone	Rest15	Rest16
Sampling	Regular	80.58	0.6123	0.4277	0.5345	0.4733	0.5901
Greedy		**79.80**	0.6157	0.4251	0.5367	0.4731	0.5912
Beam		195.69	0.6226	0.4363	0.5429	0.4787	0.5946
COT-SC	Rollback	403.22	0.6283	0.4398	0.5484	0.4802	0.5967
Ours-ER		**88.57**	0.6216	0.4382	0.5496	0.4753	0.5889
Ours-QR		143.13	0.6234	0.4420	0.5516	0.4810	0.5942
Ours-NR		164.93	0.6325	0.4397	0.5535	0.4977	0.6019
Ours-SR		104.02	**0.6382**	**0.4443**	**0.5543**	**0.5024**	**0.6043**

Table 3: Comparison of inference strategies, the speed is measured with seconds of generating 100 samples.

generate sentiment quadruples from scratch, potentially allowing for more flexibility and creativity in their outputs. Besides, we also have LLMs include closed-source zero-shot ChatGPT (Ouyang et al., 2022) and fine-tuned LLaMA-2-7B (Touvron et al., 2023) as our baselines.

As shown in Table 1 and 2, we find that generative models outperform previous classification-based methods and the structural generative method surpasses non-structural methods, this indicates that semantic structure does contribute to quadruple extraction. It also shows that the unified generation architecture can fully utilize the rich label semantics by encoding the natural language label into the target output, and it is very helpful for extracting sentiment elements jointly.

Moreover, our proposed model exhibits significant improvements over all prior studies ($p < 0.05$), demonstrating the efficacy of our rollback inference framework when applied to large language models for sentiment element generation. To the best of our knowledge, this is the first attempt to leverage semantic relations explicitly during the inference process.

5.3 Comparison of Inference Strategies

Table 3 compares the performance and computational efficiency of various inference strategies. The first three strategies follow the conventional inference approach, generating tokens forward until the end of the sequence is reached. Sampling selects the next token based on the output probability, Greedy chooses the token with the highest probability, and Beam represents beam search, could be considered as another way of generating diverse candidates. The next five strategies incorporate rollback inference, we also include COT-SC (Wang et al., 2023) as a baseline, where the rollback span covers the entire target sequence.

As evident from the results, the limited choices offered by Sampling and Greedy lead to their relatively poor performance. Beam search and COT-SC, on the other hand, improve upon these methods by maintaining a set of candidate sequences at each step. However, this comes at the cost of reduced inference speed as they must evaluate multiple candidates at each step.

Within our rollback framework, the Element Rollback inference strategy stands out for its high speed. By limiting the rollback span to individual sentiment elements, it achieves a speed close to that of Greedy inference while still leveraging contextual information for improved accuracy. Finally, if we take both aspects into consideration, the Structural Rollback inference strategy emerges as the clear winner. It outperforms all other strategies, including COT-SC, while maintaining an acceptable inference speed. We attribute this superior performance to the strategy's ability to exploit structural self-consistency associations between sentiment elements, leading to more accurate and consistent predictions.

Furthermore, case studies in Appendix A are given to make more intuitive comparisons.

6 Analysis and Discussion

In this section, we give some analysis and discussion about the robustness and effects of our rollback inference strategies.

6.1 Robustness of Rollback Inference

We first investigate if out rollback inference is robust to language models, including LLaMA-2-7B, T5-Base, and BART-Base. For each model, we evaluate both the Greedy search and Structural Rollback for a comprehensive comparison.

As shown in Table 4, our Structural Rollback inference strategy proves to be effective across all language models, consistently outperforming the greedy algorithm. This suggests that our strategy is robust and can successfully capture the associations between sentiment elements during the inference

Model	Method	Rest	Laptop	Phone	Rest15	Rest16
LLaMA	Greedy	0.6157	0.4251	0.5367	0.4731	0.5912
LLaMA	SR	**0.6382**	**0.4443**	**0.5543**	**0.5024**	**0.6043**
T5	Greedy	0.6027	0.4129	0.5246	0.4687	0.5831
T5	SR	**0.6209**	**0.4389**	**0.5489**	**0.4838**	**0.5906**
BART	Greedy	0.3956	0.3191	0.3707	0.3218	0.3893
BART	SR	**0.4177**	**0.3359**	**0.3911**	**0.3295**	**0.4042**

Table 4: Results of different language models. Rest is short for Restaurant.

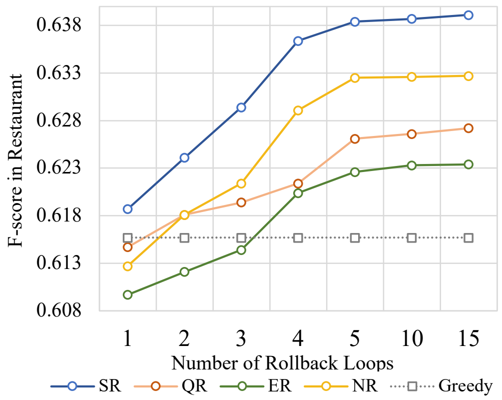

Figure 7: Performance of rollback strategies with different numbers of rollback loop.

stage, regardless of the underlying language model. This is a crucial finding as it highlights the versatility and applicability of our approach to different language models and scenarios.

Furthermore, we also investigate if our Structural Rollback is robust to the hyperparameters of generation in the Appendix B.

6.2 Impact of Rollback Loops

We further assess the impact of rollback loops on our rollback strategies. Specifically, we evaluated the performance of our rollback inference strategies in the Restaurant domain, gradually increasing the number of rollback loops from 1 to 15.

As shown in Figure 7, the performance of all our strategies consistently improved as the number of rollback loops increased, gradual leveling off after 5 and the loops more then it have very limited increase but encounter huge computational cost. This trend indicates that expanding the pool of candidates through additional rollback iterations enhances the self-consistency of large language models, leading to improved overall performance.

Among the tested strategies, Structural Rollback inference consistently outperformed the others across all loop counts, aligning with our previous experimental findings. Notably, it was the only strategy capable of surpassing greedy search even with the initial loop count of 1. This finding validates our hypothesis that leveraging the correlations among sentiment elements during inference can provide additional benefits.

Threshold	Avg. Frequency	Avg. F1	Avg. Time
0.2	0.226	0.5486	131.84
0.4	0.174	0.5483	112.38
0.6	0.151	**0.5487**	104.02
0.8	0.093	0.5424	97.16
1.0	0.042	0.5359	**89.53**

Table 5: Comparison of rollback frequency, the average frequency is calculated by average times of rollback occurred per sample.

6.3 Impact of Rollback Frequency

We subsequently investigate the impact of rollback frequency on our rollback strategies. Specifically, we adjust the rollback frequency in Structural Rollback by setting different entropy thresholds, smaller thresholds represent more rollbacks. The performance is the average of all 5 domains.

As shown in Figure 5, the performance of SR gradually grow with the increase of rollback frequency, showing that rollback does contribute to the extraction and the model's self-consistency helps mitigate issues related to local optimality that commonly afflict greedy decoding. Conversely, setting the threshold below 0.6 does not lead to further performance enhancements; Instead, it incurs a substantial computational cost. This is because the model becomes confident in its choices, resulting in repeated rollbacks to the same selections.

7 Conclusion

In this study, we move our sight to the inference process of generative ABSA and are motivated to utilize the correlations between sentiment elements during it. We thus propose a self-consistency framework named Rollback Inference Framework with a set of rollback strategies designed based on the intrinsic characteristics of the connections between sentiment elements. Experimental results show that, without requiring complex and expensive training of LLMs, our proposed inference method can achieve state-of-the-art performance in ABSA on the trade of a tiny cost in inference time.

The results also validate that, for tasks that contain semantic connections like ABSA, ignoring utilizing semantic connections during inference could lead to a waste of them.

Limitations

The limitations of our work can be stated from two perspectives. First, While our focus is on rollback inference in ABSA, it would be beneficial to explore other tasks that are closely related to ABSA. For example, event extraction, which involves identifying and extracting events from text, shares some similarities with ABSA.

Secondly, there is potential for further investigation into both unsupervised and supervised methods. Expanding the range of methods used for judging the rollback span can provide valuable insights into the strengths and weaknesses of different approaches. Supervised methods, for instance, could involve training a classifier to predict the rollback span based on labeled data, which may yield more accurate results in certain scenarios.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169.

Xiaoyi Bao, Xiaotong Jiang, Zhongqing Wang, Yue Zhang, and Guodong Zhou. 2023a. Opinion tree parsing for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7971–7984. Association for Computational Linguistics.

Xiaoyi Bao, Zhongqing Wang, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4044–4050. ijcai.org.

Xiaoyi Bao, Zhongqing Wang, and Guodong Zhou. 2023b. Exploring graph pre-training for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3623–3634, Singapore. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022. Discrete opinion tree induction for aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2064, Dublin, Ireland. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

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*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

Mengting Hu, Yinhao Bai, Yike Wu, Zhen Zhang, Liqi Zhang, Hang Gao, Shiwan Zhao, and Minlie Huang. 2023. Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13481–13494, Toronto, Canada. Association for Computational Linguistics.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2019. CAN: Constrained attention networks for multi-aspect sentiment analysis. 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 4601–4610, Hong Kong, China. Association for Computational Linguistics.

Xuancheng Huang, Zijun Liu, Peng Li, Tao Li, Maosong Sun, and Yang Liu. 2023. An extensible plug-and-play method for multi-aspect controllable text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15233–15256, Toronto, Canada. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Junjie Li, Jianfei Yu, and Rui Xia. 2022. Generative cross-domain data augmentation for aspect and opinion co-extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4219–4229, Seattle, United States. Association for Computational Linguistics.

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. 2021. Solving aspect category sentiment analysis as a text generation task. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4406–4416, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2Path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough. 2021. Open aspect target sentiment classification with natural language prompts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6311–6322, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In *COLING 2016*, pages 3298–3307.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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 2020*, pages 9122–9129.

Qianlong Wang, Zhiyuan Wen, Qin Zhao, Min Yang, and Ruifeng Xu. 2021. Progressive self-training with discriminator for aspect term extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 257–268, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. 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 2416–2429, Online. Association for Computational Linguistics.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3540–3549, Online. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Zheng Zhang, Zili Zhou, and Yanna Wang. 2022. SSEGCN: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4916–4925, Seattle, United States. Association for Computational Linguistics.

Junxian Zhou, Haiqin Yang, Yuxuan He, Hao Mou, and Junbo Yang. 2023. A unified one-step solution for aspect sentiment quad prediction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12249–12265. Association for Computational Linguistics.

A Case Study

We launch case studies to make a more intuitive comparison between our SR inference strategy and the regular Greedy generation of fine-tuned LLaMA-2-7B. We select reviews that are predicted wrongly by Greedy but have been correct through the majority vote of the candidates pool built by SR. The output formation is linearized opinion tree, the quadruples in which are organized as *(Aspect, Category, [Opinion, Polarity])*. As demonstrated in Table 6, these cases are shown in the formation of Greedy output and SR candidates pool, the majority vote would be with a ✓ notation.

The first example: Greedy gives a very typical wrong prediction, it maps "*balcony*" to "*NULL*", neglecting the adjectives "*nice*" that express clear polarity, while our method operating over majority vote, easily gives a right answer.

The second example: Greedy predicts "*friendly*" as the opinion, which is a common adjective yet not an opinion in the review since it was used to describe the unrelated content, leading to the misjudgment of sentiment polarity. Our method rollbacks the span of the sub-tree " *[friendly, Positive]*" to a right opinion and the polarity that has a strong semantic connection with it.

The third example: The root uncertain element of the Greedy sequence is "*place*", thus our SR rollbacks the entire sub-tree rooted at "*place*", which is also the entire quadruple sequence, and gets the correct output on the basis of new sub-trees with semantic connection inside them.

The fourth example: Greedy misunderstands that the "*friendly*" is used to reinforce the negative sentiment of annoying while SR salvages it with 5 loops of rollback.

The fifth example: Based on the entropy threshold, the "mercedes restaurant" is judged uncertain, thus the entire quadruple span would be our rollback span, and the majority vote gives the right answer.

From the cases shown in Table 6, we can find that, with the utilisation of the connection during inference, our method shows significant superiority in improving fine-tuned language models with a tiny cost.

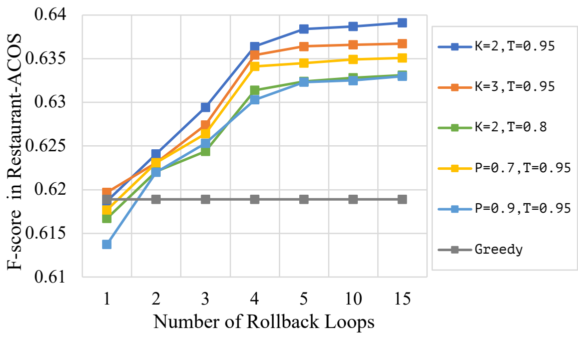

Figure 8: The performance of SR with various generation hyperparameters.

B Robustness to Hyperparameters of Generation

We further investigate the robustness of our proposed Structural Rollback towards generation hyperparameters on Restaurant-ACOS.

We show our proposed structural decoding is robust to sampling hyperparameters by varying T in temperature sampling (Ackley et al., 1985; Ficler and Goldberg, 2017), K in top-k sampling(Radford et al., 2019; Holtzman et al., 2018; Fan et al., 2018),P in nucleus sampling (Holtzman et al., 2020) in Figure 8. That gives us an conclusion that the proposed SR is robust to generation hyperparameters. Among which, we observe that the hyperparameters designed to enhance the diversity of generated content, for example, increasing K from 2 to 3, decreasing T from 0.95 to 0.8, do not contribute to the performance, we believe that is due to those strategies' purpose of increasing the diversity, will decrease the self-consistency of rollback loops.

Review text	**Method**	**Output**
if it ’ s nice outside, request for a table in the balcony	Greedy	(balcony, Ambience General, [NULL, Positive])✗
	SR Candidates Pool	(balcony, Ambience General, [nice, Positive]) ✓
		(balcony, Ambience General, [NULL, Positive]) ✗
		(balcony, Ambience General, [nice, Positive]) ✓
		(balcony, Ambience General, [nice, Positive]) ✓
		(balcony, Ambience General, [NULL, Positive]) ✗
the prior reviews said kid friendly	Greedy	(NULL, Restaurant Miscellaneous, [friendly, Positive])✗
	SR Candidates Pool	(NULL, Restaurant Miscellaneous, [friendly, Positive]) ✗
		(NULL, Restaurant Miscellaneous, [friendly, Positive]) ✗
		(NULL, Restaurant Miscellaneous, [NULL, Negative]) ✓
		(NULL, Restaurant Miscellaneous, [NULL, Negative]) ✓
		(NULL, Restaurant Miscellaneous, [NULL, Negative]) ✓
i highly recommend this place to all that want to try indain food for the first time	Greedy	(place, Restaurant Miscellaneous, [recommend, Positive])✗
	SR Candidates Pool	(indain food, Food Quality, [recommend, Positive]) ✓
		(indain food, Food Quality, [recommend, Positive]) ✓
		(indain food, Food Quality, [recommend, Positive]) ✓
		(indain food, Food Quality, [recommend, Positive]) ✓
		(place, Restaurant Miscellaneous, [recommend, Positive]) ✗
but she is very friendly with certain people , making it even more annoying	Greedy	(NULL, Service General, [friendly, Negative])✗
	SR Candidates Pool	(NULL, Service General, [friendly, Negative]) ✗
		(NULL, Service General, [annoying, Negative]) ✓
		(NULL, Service General, [annoying, Negative]) ✓
		(NULL, Service General, [annoying, Negative]) ✓
		(NULL, Service General, [friendly, Negative]) ✗
mercedes restaurant is so tasty, the service is undeniably awesome	Greedy	(mercedes restaurant, Food Quality, [tasty, Positive])✗
	SR Candidates Pool	(mercedes restaurant, Food Quality, [tasty, Positive]) ✗
		(NULL, Food Quality, [tasty, Positive]) ✓
		(mercedes restaurant, Food Quality, [tasty, Positive]) ✗
		(NULL, Food Quality, [tasty, Positive]) ✓
		(NULL, Food Quality, [tasty, Positive]) ✓

Table 6: Cases study, the quadruples in which are organized in *(Aspect, Category, [Opinion, Polarity])* as introduced in Figure 5.