









Domain-Expanded ASTE: Rethinking Generalization in Aspect Sentiment Triplet Extraction

Yew Ken Chia^{* 1, } Hui Chen^{} Guizhen Chen^{1, 2 } Wei Han^{}
Sharifah Mahani Aljunied^{1 } Soujanya Poria^{} Lidong Bing^{1 }

 Singapore University of Technology and Design

¹DAMO Academy, Alibaba Group, Singapore

²Nanyang Technological University, Singapore

sporia@sutd.edu.sg guizhen001@ntu.edu.sg

{yewken_chia, hui_chen, wei_han}@mymail.sutd.edu.sg

{yewken.chia, guizhen.chen, mahani.aljunied, l.bing}@alibaba-inc.com

Abstract

Aspect Sentiment Triplet Extraction (ASTE) is a challenging task in sentiment analysis, aiming to provide fine-grained insights into human sentiments. However, existing benchmarks are limited to two domains and do not evaluate model performance on unseen domains, raising concerns about the generalization of proposed methods. Furthermore, it remains unclear if large language models (LLMs) can effectively handle complex sentiment tasks like ASTE. In this work, we address the issue of generalization in ASTE from both a benchmarking and modeling perspective. We introduce a domain-expanded benchmark by annotating samples from diverse domains, enabling evaluation of models in both in-domain and out-of-domain settings. Additionally, we propose CASE, a simple and effective decoding strategy that enhances trustworthiness and performance of LLMs in ASTE. Through comprehensive experiments involving multiple tasks, settings, and models, we demonstrate that CASE can serve as a general decoding strategy for complex sentiment tasks. By expanding the scope of evaluation and providing a more reliable decoding strategy, we aim to inspire the research community to reevaluate the generalizability of benchmarks and models for ASTE. Our code, data, and models are available at <https://github.com/DAMO-NLP-SG/domain-expanded-aste>.

1 Introduction

Opinions and sentiments are essential to human communication, beliefs, and behaviors (Liu, 2012). Although sentiment analysis is often performed

^{*}Yew Ken and Guizhen are students under the Joint PhD Program between Alibaba and their corresponding university. This work is done by Hui and Wei during internship at Alibaba.

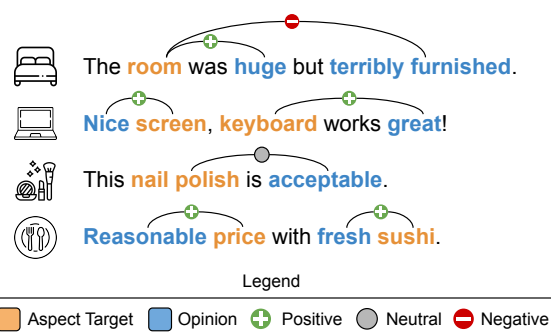


Figure 1: ASTE data samples for the Hotel, Laptop, Cosmetics, and Restaurant domains, respectively.

at the sentence or document level, it is insufficient to capture the fine-grained sentiment information and nuances of human opinions (Poria et al., 2020). To this end, aspect sentiment triplet extraction (ASTE) (Peng et al., 2020) is a challenging and well-established task of aspect-based sentiment analysis (Pontiki et al., 2014) which aims to extract richer and more interpretable sentiment information from natural language. Concretely, ASTE considers how each opinion term in a text may express sentiments towards specific aspect targets.

Although ASTE has become a more established task with many existing methods (Zhang et al., 2022), we are concerned that they may not generalize well due to limitations in the existing benchmark datasets. Notably, the established benchmarks are limited to two domains, which limits the evaluation scope of model capabilities and does not represent the diversity of real-world data. On the other hand, it is also important to assess how models generalize to unseen domains as domain-specific labeled data is often scarce (Wang and Pan, 2018), and models may face domain shift during deployment (Wang et al., 2021). Hence, this mo-

tivates us to propose a domain-expanded ASTE benchmark which not only considers the in-domain performance, but also evaluates out-of-domain generalization across a more diverse set of domains. We support the new benchmark by annotating more than 4,000 data samples for two new domains based on hotel and cosmetics product reviews. Therefore, we can construct a domain-expanded dataset with four domains as shown in Figure 1.

To investigate the domain generalization of existing ASTE methods, we evaluate five existing methods based on pretrained language models (PLMs) for the in-domain and out-of-domain settings. On the other hand, while large language models (LLMs) have recently enabled breakthroughs in many NLP tasks, it is unclear if they can surpass specialized pretrained language models (PLMs) on sentiment tasks such as ASTE (Zhang et al., 2023). Despite the impressive language understanding and general-purpose capabilities of LLMs, it is challenging to adapt them to ASTE due to several reasons. Notably, black-box models like GPT-4 (OpenAI, 2023) are less trustworthy and interpretable as it is not clear how to estimate the confidence of their predictions. For instance, as each text may contain multiple sentiment triplets, it is useful to know which of the predicted triplets have higher confidence or lower confidence. Hence, the lack of interpretability hinders the trustworthiness of LLMs in practical applications, and limits in-depth analysis of their performance. On the other hand, it is generally not possible or feasible to train LLMs for specific tasks, leading to greater focus on prompt-based methods to improve performance.

Thus, we introduce confidence-aware sentiment extraction (CASE), a simple and effective decoding strategy to improve the trustworthiness and performance of LLMs for complex sentiment tasks like ASTE. Inspired by self-consistency (Wang et al., 2023a) which samples diverse reasoning paths to select the most consistent answer, we sample diverse sets of sentiment triplets to select the most consistent triplets. Intuitively, sentiment triplets which are most consistent, i.e., occur most often when sampling diverse sets of triplets, can be assigned a higher confidence. Notably, it is simple to integrate CASE with any language model that supports stochastic sampling, and it does not require any model re-training or access to model logits. Compared to conventional decoding methods such as greedy search or beam search, CASE en-

hances interpretability by estimating the confidence of each predicted triplet, and improves performance by explicitly considering a larger pool of sentiment triplets.

In summary, our main contributions include: (1) To evaluate ASTE methods more holistically, we propose a domain-expanded benchmark which covers in-domain and out-of-domain performance across diverse domains. (2) We annotate more than 4000 samples for two new domains based on hotel and cosmetics product reviews to support the new benchmark. (3) We propose CASE, a simple and effective decoding strategy to enhance the trustworthiness and performance of LLMs for ASTE. Our experiments demonstrate its effectiveness across different models, tasks, and settings.

2 Related Work

Aspect-Based Sentiment Analysis While sentiment analysis is often considered at the sentence or document level, this approach cannot capture the fine-grained sentiment information and nuances of human opinions (Poria et al., 2020). To this end, aspect-based sentiment analysis (ABSA) consists of many tasks which aim to reveal richer sentiment information by considering the specific opinions and aspect targets in natural language (Pontiki et al., 2014). Early works on ABSA focused on extracting individual sentiment elements, such as aspect term extraction (Liu et al., 2015), opinion term extraction (Yang and Cardie, 2012), or aspect sentiment classification (Dong et al., 2014). On the other hand, compound ABSA tasks have been introduced to jointly address multiple subtasks, including ASTE (Peng et al., 2020) and ASQP (Zhang et al., 2021a). In this work, we focus on ASTE which has many established methods, yet has not been studied through the lens of domain generalization (Wang et al., 2021).

Domain Generalization While traditional machine learning methods are trained based on the assumption that training and testing data are identically and independently distributed, this assumption seldom holds true in reality. Hence, the performance of methods often deteriorates due to shifts in domain distributions (Wang et al., 2021). As it is not feasible to comprehensively annotate task-specific data for training, there is an urgent need to improve the robustness and generalization ability of existing methods. While there are many related topics such as domain adaptation (Patel et al.,

2015; Gong et al., 2020), meta-learning (Vilalta and Drissi, 2002), and lifelong learning (Biesialska et al., 2020), we believe that domain generalization is more widely applicable to the established methods for ASTE. Hence, in this work, we mainly investigate domain generalization, the goal of which is to learn a model that will generalize well to unseen domains.

Large Language Models Recently, there have been numerous advancements in natural language processing due to the rapid development of large language models (LLMs) such as GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023). Compared to the smaller pretrained language models (PLMs), LLMs have deeper language understanding and reasoning capabilities, owing to the large scale of the models and training data. Moreover, the performance of LLMs can be further enhanced through methods such as instruction-tuning (Wei et al., 2022a), chain-of-thought prompting (Wei et al., 2022c), and reinforcement learning from human feedback (Ouyang et al., 2022). However, there is less focus on fundamental decoding strategies that can heavily affect the behavior of generative methods. On the other hand, language models are prone to hallucinating outputs that seem plausible but are incorrect or unreasonable (Ji et al., 2022), raising major concerns about their trustworthiness and interpretability (Zhao et al., 2023). Hence, we introduce a novel decoding strategy that aims to improve the performance and interpretability of LLMs for ASTE.

3 Domain-Expanded ASTE Benchmark

To evaluate the performance of ASTE methods more holistically and encourage development of more robust methods, we propose a domain-expanded benchmark. The benchmark assesses models not only in-domain, but also in terms of out-of-domain generalization across diverse domains. Hence, we construct the benchmark by leveraging two domains from existing datasets, while annotating samples for two new domains. In this section, we detail the dataset construction process and dataset statistics for each domain.

3.1 Task Formulation

Given an input sentence x containing n words, ASTE aims to predict a set of sentiment triplets where each triplet (t, o, p) corresponds to the aspect target, opinion, and sentiment polarity, respectively.

Domain	Aspect Target	Opinion	Sentiment Triplet
Hotel	0.73	0.76	0.61
Cosmetics	0.72	0.73	0.57

Table 1: Inter-annotator agreement scores. We measure the agreement using the AvgAgr metric separately for aspect targets, opinions, and sentiment triplets.

Each aspect target t and opinion o are text spans in the sentence. The sentiment polarity belongs to the label set of {POS, NEG, NEU}, which corresponds to positive, negative, and neutral sentiment, respectively.

3.2 Data Collection

We construct a dataset with four domains by leveraging two domains from existing datasets (Peng et al., 2020) and collecting data for two new domains. Specifically, we collect review texts in the Hotel and Cosmetics domains from TripAdvisor Reviews (Angelidis et al., 2021) and Amazon Reviews (He and McAuley, 2016; McAuley et al., 2015) respectively. We collect 8000 samples from each domain corpus and use the spaCy tool to tokenize the review texts and label their part-of-speech tags. To denoise the raw samples, we remove reviews that do not contain any nouns or adjectives. We also leverage the existing Laptop and Restaurant domains from ASTE-Data-V2 (Xu et al., 2020). Within the Laptop and Restaurant domains, we remove duplicate samples and retain the existing triplet annotations.

Domain	#Train	#Dev	#Test	#Triplets	#T	#O
Restaurant	1771	442	739	5376	1878	1743
Laptop	867	217	362	2334	1086	1083
Hotel	1281	320	535	4064	1486	1706
Cosmetics	1287	442	739	4002	1539	2221

Table 2: Statistics of our domain-expanded ASTE dataset. We report the number of train samples, development samples, test samples, sentiment triplets, unique aspect targets (T), and unique opinions (O).

3.3 Data Annotation

For annotation, we follow the same data format as existing datasets (Peng et al., 2020; Xu et al., 2020). Specifically, annotators are provided with each tokenized review sentence as input. They are required to annotate all valid sentiment triplets in the text according to the task formulation in Section 3.1. We include the detailed annotation

guideline in the appendix. To ensure the quality of data annotation, we conduct quality checking for each batch of annotated data. Specifically, for each annotation batch, 10% of the samples are randomly selected for manual checking. If more than 10% of the selected samples contain errors, we provide detailed feedback and request annotators to amend the batch. We engage two independent annotators to label the data and engage a third annotator to resolve any annotation disagreements.

Following previous works in data annotation for ABSA (Barnes et al., 2018), we measure the inter-annotator agreement using the AvgAgr metric (Wiebe et al., 2005):

$$\text{AvgAgr}(a, b) = \frac{1}{2} \left(\frac{|a \cap b|}{|a|} + \frac{|a \cap b|}{|b|} \right) \quad (1)$$

where a and b are the set of annotations by the first and second annotators, respectively. Intuitively, the agreement value is the average of precision and recall between the two annotators. Hence, the perfect agreement is 1 while no agreement is 0. We report the inter-annotator agreement for the Hotel and Cosmetics domain in Table 1. We observe that the agreement scores are high and comparable to previous ABSA datasets (Barnes et al., 2018).

We report the statistics¹ of the domain-expanded dataset such as the number of reviews, sentiment triplets, and unique aspect targets in Table 2.

4 Confidence-Aware Sentiment Extraction (CASE)

To enhance the trustworthiness and effectiveness of large language models (LLMs) on ASTE, we propose confidence-aware sentiment extraction (CASE), a simple and effective decoding strategy. Compared to conventional decoding methods such as greedy search or beam search, CASE enhances interpretability by estimating the confidence of each predicted triplet, and improves performance by explicitly considering a larger pool of sentiment triplets. Inspired by self-consistency (Wang et al., 2023a) which samples diverse reasoning paths to select the most consistent answer, we sample diverse sets of sentiment triplets to select the most consistent triplets. As shown in Figure 2, CASE consists of four main steps: (1) Given the input text, we sample diverse output sequences from the language model, where each output sequence represents a set of candidate sentiment triplets. (2) The

unique sentiment triplets are then aggregated based on the sampled sets of triplets. (3) To estimate the confidence of each sentiment triplet, we calculate the occurrence frequency of each triplet. (4) Lastly, we select the most confident sentiment triplets as the final predictions.

4.1 Candidate Sampling

In practice, generative methods such as sequence-to-sequence PLMs (Zhang et al., 2021a,b) and LLMs (Wang et al., 2023b; Zhang et al., 2023) use approximate decoding methods such as greedy search or beam search as it is intractable to determine the optimal y for a given input x , i.e., $\text{argmax}_y p(y | x)$. Hence, we argue that generating a single sequence y is sub-optimal as it only provides a narrow view of the possible triplet candidates. On the other hand, sampling diverse sequences from the language model can provide the opportunity to consider a larger set of triplet candidates and estimate the confidence score of each triplet. To obtain diverse triplet candidates, we use temperature-based sampling (Ficler and Goldberg, 2017; Fan et al., 2018) which is a common method to generate diverse outputs from a language model. Concretely, we sample m outputs from our model G for a given input x :

$$S_j \sim G(x, k), j \in \{1, \dots, m\} \quad (2)$$

where S_j denotes the set of sentiment triplets in the j -th sampled sequence.

4.2 Aggregation

Naturally, a triplet set may be sampled more than once and a sentiment triplet (t, o, p) may be present in more than one set. To aggregate the sentiment triplets, we take the union of the sampled sets to form the candidate set S_c :

$$S_c = \bigcup_{j=1}^m S_j \quad (3)$$

Hence, we only consider the unique sentiment triplets across all the sampled triplet sets.

4.3 Confidence Estimation

Intuitively, we assume that sentiment triplets that appear more frequently can be attributed to a higher confidence score. Thus, we estimate the confidence score of each sentiment triplet $(t, o, p) \in S_c$ to be

¹We include more detailed analysis in Appendix A.7.

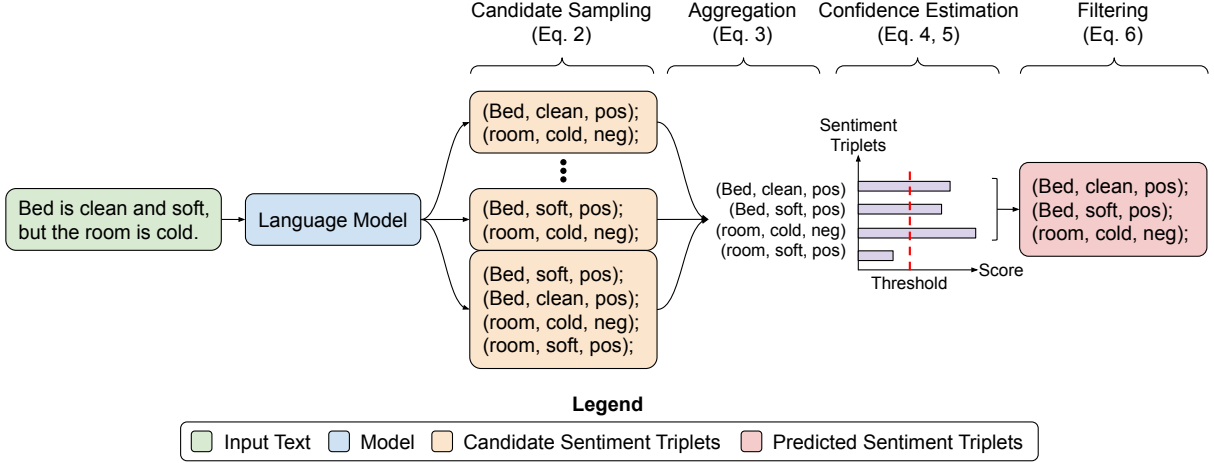


Figure 2: Our proposed confidence-aware sentiment extraction (CASE) decoding strategy which aims to enhance the trustworthiness and performance of LLMs for ASTE.

the corresponding occurrence frequency:

$$\phi(t, o, p) = \frac{\sum_{j=1}^m \mathbf{1}_{S_j}(t, o, p)}{m} \quad (4)$$

where $\mathbf{1}_{S_j}(t, o, p)$ is the indicator function of whether a triplet (t, o, p) appears in S_j :

$$\mathbf{1}_{S_j}(t, o, p) = \begin{cases} 1 & \text{if } (t, o, p) \in S_j \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Naturally, the confidence score for each triplet is bounded within the range $0 \leq \phi(t, o, p) \leq 1$. As it is not feasible to exhaustively sample from the language model, we sample $m = 20$ output sequences for each input x .

4.4 Filtering

While the steps thus far have improved interpretability through confidence estimation and triplet recall by sampling a larger pool of candidate triplets, we face the challenge noisy predictions. Specifically, sampling more triplets may impact model precision due to increased numbers of false positive triplets. Hence, we apply a confidence threshold T over each triplet $(t, o, p) \in S_c$ to select the final prediction set S_{final} :

$$S_{\text{final}} = \{(t, o, p) \mid \phi(t, o, p) \geq T\} \quad (6)$$

This filtering process ensures that we retain only the higher-confidence triplets, thus mitigating noisy predictions.

5 Experiment Setup

5.1 Settings

In this work, we aim to provide a more holistic study of model performance on the ASTE task.

While previous works mostly focus on the in-domain setting, where the model is trained and tested on the same domain, we believe that this provides a limited perspective of model performance, as it does not consider robustness to domain shift. Hence, we further evaluate models out-of-domain settings, where the model trained on one domain and tested on a different domain. Moreover, certain models may be stronger in low-resource scenarios, which is important to consider as labeled data is often limited and costly to obtain in practice. Thus, we further assess each model on the fully-supervised and few-shot scenarios. Specifically, for the few-shot scenario, we sample 5 examples for each sentiment polarity. Following previous works in ASTE (Peng et al., 2020; Xu et al., 2020), we use the F_1 metric to measure model performance. For all training experiments, we report the average results from 5 random runs.

5.2 Models

To provide a study of diverse models, we evaluate several ASTE methods based on pretrained language models (PLMs) and large language models (LLMs). For PLMs, we include discriminative methods including GTS (Wu et al., 2020) based on sequence tagging, Span-ASTE based on span enumeration and RoBMRC (Liu et al., 2022b) based on machine reading comprehension. We also consider generative methods including GAS (Zhang et al., 2021b) and Paraphrase (Zhang et al., 2021a). As LLMs have shown general-purpose capabilities and strong performance on many language understanding and reasoning tasks, we also assess their performance on ASTE. Specifically, we use the ChatGPT

model API based on gpt-3.5-turbo-0301². We note that while LLMs are technically PLMs as they also undergo large-scale pretraining, we use PLMs to refer to smaller models that are pretrained, such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020). To adapt ChatGPT to complex sentiment tasks such as ASTE, we use in-context learning demonstrations (Wei et al., 2022b) with the prompt templates as shown in Appendix A.6. For the fully supervised scenario, we leverage in-context demonstration selection (Liu et al., 2022a) which selects relevant examples from the full dataset based on cosine similarity. Specifically, we use embedding representations from Sentence-BERT (Reimers and Gurevych, 2019) and select the top-15 most similar examples as in-context demonstrations. For the few-shot scenario, we use the few-shot examples as in-context demonstrations.

5.3 Hyperparameters

For all PLM-based methods, we use the base model size and original hyperparameters for training experiments. For sampling with CASE, we generate a fixed number of 10 outputs for each example. To select the confidence threshold hyperparameter T , we perform a grid search with the values $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ based on F_1 results on the development set. For out-of-domain settings, we choose the confidence threshold from the respective source domain. In addition, we report other experimental details in Appendix A.2.

6 Results and Analysis

To provide a holistic study of ASTE methods, we evaluate on the proposed domain-expanded ASTE benchmark, reporting the fully supervised in-domain results in Table 3, with fully supervised out-of-domain results in Table 4. We further study the few-shot scenario for in-domain and out-of-domain settings in Table 5. In general, while specialized PLM-based methods currently outperform LLMs in the fully supervised scenario, there is a smaller performance gap for unseen domains, and LLMs exhibit better robustness to domain shift. In contrast, we find that LLMs are more effective in low-resource scenarios, as evidenced by the few-shot results. On the other hand, we observe that the proposed CASE is an effective decoding strategy that not only addresses the fundamental interpretability

limitation of LLMs, but also consistently improves performance across models, settings, and tasks.

6.1 Fully Supervised Results

Evaluation of PLM-Based Methods Based on the established methods that leverage PLMs, we find significant differences in performance and generalization for generative methods (i.e., Paraphrase, GAS) compared to discriminative methods (i.e., GTS, Span-ASTE, RoBMRC). Specifically, generative methods enjoy competitive in-domain performance and much stronger generalization to unseen domains, with an advantage of more than 2 points in the out-of-domain setting on average. Furthermore, while PLM-based methods generally demonstrate large performance disparities between in-domain and out-of-domain settings, generative methods are more robust to domain shift, as they exhibit smaller performance gaps on average (14.58) compared to discriminative methods (16.80). We believe that this is largely due to the effect of label semantics (Ma et al., 2022). For instance, understanding that “fresh” is an adjective for describing food such as “sushi” in Figure 1, it can be easier for the model to predict the sentiment triplet (sushi, fresh, positive). Hence, generative methods demonstrate better performance and generalization on the domain-expanded benchmark.

Comparison of LLM-Based Methods By comparing the LLM-based ChatGPT to PLM-based methods, we observe that LLMs perform worse in general for fully-supervised scenarios, but show greater robustness to domain shift. Notably, ChatGPT performs significantly worse on in-domain settings compared to PLM-based methods for ASTE. This is in contrast to their strong performance on simpler sentiment tasks such as sentence-level sentiment classification (Zhang et al., 2023). We believe that the difficulty that LLMs face in ASTE stems from the complexity of the task, as the structured nature of the sentiment triplets are less natural for language models. Hence, there is larger area of improvement for task-specific adaptation of LLMs, especially for complex tasks such as ASTE. On the other hand, we observe that ChatGPT can attain similar out-of-domain performance compared to some PLM-based methods, with a smaller performance gap between in-domain and out-of-domain settings (7.4). We posit that the greater robustness to domain shift is due to exposure to more diverse pretraining data, which together with model

²<https://platform.openai.com/docs/models/gpt-3-5>

Method	Hotel			Laptop			Cosmetics			Restaurant			Avg.
	<i>P.</i>	<i>R.</i>	<i>F</i> ₁	<i>P.</i>	<i>R.</i>	<i>F</i> ₁	<i>P.</i>	<i>R.</i>	<i>F</i> ₁	<i>P.</i>	<i>R.</i>	<i>F</i> ₁	<i>F</i> ₁
GTS (Wu et al., 2020)	58.76	59.50	59.13	58.07	48.16	52.65	51.42	50.95	51.18	65.06	65.45	65.26	57.15
Span-ASTE (Xu et al., 2021)	67.73	62.92	65.24	60.73	54.40	57.39	59.79	55.0	57.29	68.69	65.41	67.01	61.74
RoBMRC (Liu et al., 2022b)	68.99	63.11	65.92	66.12	51.51	57.90	58.62	55.27	56.89	69.89	67.80	68.83	62.49
Paraphrase (Zhang et al., 2021a)	65.21	61.07	63.08	61.23	55.13	58.02	58.45	53.62	55.93	68.56	68.46	68.51	61.41
GAS (Zhang et al., 2021b)	67.57	63.30	65.37	60.59	55.13	57.73	59.13	55.53	57.28	69.26	69.16	69.21	62.41
with CASE (Ours)	67.40	64.75	66.05	60.60	56.79	58.63	59.51	57.01	58.23	68.84	70.42	69.62	63.13
ChatGPT	47.59	53.13	50.20	44.57	49.12	46.74	34.80	38.73	36.66	53.49	57.68	55.50	47.28
with CASE (Ours)	54.24	49.86	51.96	51.71	48.17	49.88	42.32	35.39	38.55	58.11	56.04	57.06	49.36

Table 3: Evaluation results for **in-domain** ASTE with the full datasets. We report the average precision (*P.*), recall (*R.*), and *F*₁ scores for each domain, as well as the average *F*₁ (Avg.) across all domains.

Method	Hotel			Laptop			Cosmetics			Restaurant			Avg.
	L→H	C→H	R→H	H→L	C→L	R→L	H→C	L→C	R→C	H→R	L→R	C→R	<i>F</i> ₁
GTS (Wu et al., 2020)	35.05	52.75	49.41	34.01	32.68	40.98	38.08	24.31	32.77	55.73	49.86	49.94	41.65
Span-ASTE (Xu et al., 2021)	41.62	55.55	51.23	37.34	33.48	42.52	43.55	31.00	34.30	57.31	54.36	51.44	44.58
RoBMRC (Liu et al., 2022b)	36.17	58.17	52.67	37.77	35.57	41.26	41.81	26.97	32.12	60.47	51.10	55.73	44.76
Paraphrase (Zhang et al., 2021a)	43.99	56.49	50.81	41.71	39.09	48.02	43.85	28.45	34.68	59.74	59.15	56.14	46.90
GAS (Zhang et al., 2021b)	46.18	59.10	52.71	40.77	37.88	48.25	46.10	29.81	34.97	59.57	60.47	56.54	47.76
with CASE (Ours)	46.84	60.06	53.32	42.36	38.82	48.72	47.77	30.96	36.12	60.03	61.06	57.08	48.60
ChatGPT	42.98	42.61	43.14	34.48	35.23	36.43	31.22	31.26	31.84	50.08	51.29	48.02	39.88
with CASE (Ours)	42.91	45.07	45.56	36.08	36.66	38.57	31.04	31.60	32.80	51.78	53.74	50.43	41.35

Table 4: Evaluation results for **out-of-domain** ASTE with the full datasets. We report the average *F*₁ score for each domain-pair (source domain → target domain), as well as the average *F*₁ (Avg.) across all domain-pairs.

Method	In-Domain <i>F</i> ₁	Out-Of-Domain <i>F</i> ₁
Span-ASTE	32.65	20.71
Paraphrase	33.46	22.95
GAS	36.53	26.72
with CASE (Ours)	38.42	28.81
ChatGPT	44.38	38.19
with CASE (Ours)	47.34	39.56

Table 5: Evaluation results for **few-shot** ASTE (5-Shot). We report the average in-domain *F*₁ score across all domains, and the average out-of-domain *F*₁ score across all domain-pairs.

scaling, imbues LLMs with comprehensive world knowledge (Safavi and Koutra, 2021). This is consistent with previous findings that training data diversity is the main factor in robustness to domain shift (Taori et al., 2020). Thus, LLM-based methods show promising generalization to new domains, with ample room for future development.

6.2 Few-Shot Performance

In contrast to the fully supervised results, we find that LLMs show stronger performance in low-resource scenarios, as shown in Table 5. Notably, ChatGPT significantly outperforms the PLM-based methods in both the in-domain and out-of-domain

settings. As LLMs benefit from massive scale of model parameters and training data, this enables them to learn a wider range of language patterns and semantics, hence generalizing well to new tasks, even with limited data (Brown et al., 2020). From a practical point of view, while there remains ample room for improvement in the fully supervised scenarios, the strong generalization in low-resource scenarios and robustness to domain shift make LLMs suitable for data-scarce applications. Hence, we believe that the few-shot results highlight the importance of evaluating ASTE methods on diverse scenarios, in order to provide a holistic view of their capabilities.

6.3 Impact of CASE

While our proposed CASE decoding strategy was mainly motivated by the limitations of interpretability and trustworthiness of black-box LLMs for ASTE, we find that it also provides reliable performance benefits. Notably, we observe that ChatGPT with CASE consistently outperforms the baseline which uses greedy decoding³ for both in-domain as well as out-of-domain settings. Furthermore, as our decoding strategy is applicable to any method that

³While we have also experimented with beam search, we observed similar performance and hence used greedy search.

Task	Dataset	Method	Orig.	w/ CASE
AOPE	Hotel	GAS (Zhang et al., 2021b)	71.77	72.44
	Laptop		65.93	66.77
	Cosmetics		62.98	63.91
	Restaurant		75.33	75.51
ASQP	Rest15	Paraphrase (Zhang et al., 2021a)	46.93	47.96
	Rest16		57.93	58.86

Table 6: Evaluation results for **in-domain** ABSA subtasks when using generative methods without change or with confidence-aware generative extraction (CASE).

supports stochastic sampling, we easily apply it to the generative method GAS (Zhang et al., 2021b), which also shows consistent benefits. We believe that the performance benefits of CASE stem mainly from the sampling process which considers more diverse sentiment triplets, which is supported by the significantly improved recall scores in Table 3. On the other hand, there is little to no negative impact on precision, which suggests that our aggregation and filtering steps can effectively mitigate false positive triplets. This is in contrast to conventional decoding methods such as greedy decoding, which only presents a single, less optimal set of sentiment triplets for consideration. Hence, we believe that CASE is an effective decoding strategy for ASTE and a promising direction for future development.

6.4 Benefit of CASE on Other ABSA Tasks

As CASE is a decoding strategy that can enhance the performance of generative models, it may also benefit other ABSA tasks. Hence, to further study its effectiveness, we report the in-domain results of CASE-based generative models for aspect opinion pair extraction (AOPE) (Chen et al., 2020) and aspect sentiment quad prediction (ASQP) (Zhang et al., 2021a). We use our domain-expanded dataset for AOPE and the original Rest15 and Rest16 datasets for ASQP. To modify our method for AOPE and ASQP, we simply consider pair sets and quadruplet sets respectively in the sampling process instead of triplet sets for ASTE. Note that our method does not affect model parameters or re-training any models to be re-trained. Based on the results in Table 6, we observe consistent improvement when using generative methods with CASE compared to using the original greedy decoding. Furthermore, it can improve the interpretability and trustworthiness of generative ABSA predictions by estimating the confidence score of each pair, triplet, or quadruplet. Overall, we believe that CASE can be a beneficial and widely applicable technique for different ABSA tasks.

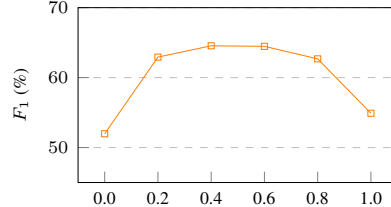


Figure 3: The effect of confidence-aware threshold T on in-domain performance for the Hotel domain.

6.5 Effect of Confidence-Aware Threshold

As CASE aims to improve the model recall while reducing false positives, it is crucial to remove the low-confidence triplets by applying a sufficiently high threshold filter. However, a threshold that is too high may introduce more false negative triplets. Hence, we study the effect of the confidence-aware threshold T on model performance in Figure 3. We find that the in-domain performance is relatively stable across a wide range of thresholds between 0.2 and 0.8. This suggests that the false positive triplets mainly have very low confidence scores i.e., $\phi(t, o, p) < 0.2$. However, there is a sharp decrease in performance for extremely low or high threshold values, which is consistent with our intuition.

7 Conclusions

In conclusion, this work addressed the task of Aspect Sentiment Triplet Extraction (ASTE) in sentiment analysis, focusing on the issues of limited benchmark domains and the challenges of large language models (LLMs) in handling complex sentiment tasks. We introduced a domain-expanded ASTE benchmark by annotating samples from diverse domains, enabling the evaluation of models in both in-domain and out-of-domain settings. This expanded benchmark provided a more comprehensive assessment of model performance, addressing concerns regarding the generalizability of proposed methods. Secondly, a novel decoding strategy called CASE (Context-Aware Sampling and Enhancement) was proposed to enhance the trustworthiness and performance of LLMs in ASTE.

The experimental results demonstrated its effectiveness across multiple tasks, settings, and models. Its simplicity and efficacy make it a promising general decoding strategy for complex sentiment tasks. By expanding the scope of evaluation and providing a reliable decoding strategy, we hope to encourage the research community to rethink the generalizability of benchmarks and models for ASTE. The findings highlight the importance of considering diverse domains and utilizing appropriate decoding strategies when tackling fine-grained sentiment analysis tasks. With these contributions, we hope to foster the development of more robust and capable sentiment analysis methods in the future.

Acknowledgment

This work was substantially supported by DAMO Academy through DAMO Academy Research Intern Program.

Limitations

As our method samples multiple output sequences for a given input sequence, there is an increased computational cost for inference. However, this is a trade-off similar to tuning hyperparameters for beam search in text generation problems, and the effect can be mitigated by batched inference. Our method also relies on the sampled sequences to have sufficient diversity in order to consider a larger set of candidate triplets. However, too much diversity may introduce unwanted noise. The diversity is affected by both the temperature sampling hyperparameter and the number of sampled sequences. In this work, we keep the temperature sampling hyperparameter fixed as a standard value for generation due to computational constraints. We analyze the effect of the number of sampled sequences m in Appendix A.3.

Ethics Statement

For data annotation, we engage two professional annotators who are fairly compensated. The compensation is negotiated based on the task complexity and assessment of a reasonable annotation speed. The annotators have given their consent for their annotations to be publicly released as a research dataset. The data annotation project passes the ethics review of the data annotation team as it does not contain any confidential data. The data annotators are adults who are versed in multiple languages. We release our datasets under the same license (CC

BY NC 4.0) as the original data that we collected from. The license allows for free sharing and adaptation of the dataset as long as appropriate credit is given, and the data is only used for non-commercial purposes.

References

- Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. [Extractive opinion summarization in quantized transformer spaces](#). *Transactions of the Association for Computational Linguistics*, 9:277–293.
- Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. [MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Magdalena Biesialska, Katarzyna Biesialska, and Marta Ruiz Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. *ArXiv*, abs/2012.09823.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. [Synchronous double-channel recurrent network for aspect-opinion pair extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6515–6524, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. [Adaptive recursive neural](#)

- network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. 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 Fidler 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.
- Chengcong Gong, Jianfei Yu, and Rui Xia. 2020. Unified feature and instance based domain adaptation for aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7035–7045, Online. Association for Computational Linguistics.
- Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1 – 38.
- Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Lisbon, Portugal. Association for Computational Linguistics.
- Shu Liu, Kaiwen Li, and Zuhe Li. 2022b. A robustly optimized BMRC for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 272–278, Seattle, United States. Association for Computational Linguistics.
- Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. Label semantics for few shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1956–1971, Dublin, Ireland. Association for Computational Linguistics.
- Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.
- OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.
- 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 E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.
- Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. 2015. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69.
- Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8600–8607.
- Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *ArXiv*, abs/2005.00357.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

- and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Tara Safavi and Danai Koutra. 2021. **Relational World Knowledge Representation in Contextual Language Models: A Review**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1053–1067, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Rohan Taori, Achal Dave, Vaishal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. **Measuring robustness to natural distribution shifts in image classification**. In *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599. Curran Associates, Inc.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. **Llama: Open and efficient foundation language models**. *ArXiv*, abs/2302.13971.
- Ricardo Vilalta and Youssef Drissi. 2002. A perspective view and survey of meta-learning. *Artificial Intelligence Review*, 18:77–95.
- Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. 2021. **Generalizing to unseen domains: A survey on domain generalization**. *ArXiv*, abs/2103.03097.
- Wenya Wang and Sinno Jialin Pan. 2018. **Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2171–2181, Melbourne, Australia. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. **Self-consistency improves chain of thought reasoning in language models**. In *The Eleventh International Conference on Learning Representations*.
- Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023b. **Is chatgpt a good sentiment analyzer? a preliminary study**. *ArXiv*, abs/2304.04339.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. **Finetuned language models are zero-shot learners**. In *International Conference on Learning Representations*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. **Emergent abilities of large language models**. *Transactions on Machine Learning Research*. Survey Certification.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022c. **Chain of thought prompting elicits reasoning in large language models**. *ArXiv*, abs/2201.11903.
- Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. **Annotating expressions of opinions and emotions in language**. *Language Resources and Evaluation*, 39:165–210.
- Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. **Grid tagging scheme for aspect-oriented fine-grained opinion extraction**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.
- Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. **Learning span-level interactions for aspect sentiment triplet extraction**. 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 4755–4766, Online. Association for Computational Linguistics.
- 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.
- Bishan Yang and Claire Cardie. 2012. **Extracting opinion expressions with semi-Markov conditional random fields**. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1335–1345, Jeju Island, Korea. 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, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. **Sentiment analysis in the era of large language models: A reality check**.
- 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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *ArXiv*, abs/2203.01054.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

A Appendix

A.1 Duplicate-Aware Evaluation for ASTE

Algorithm 1: Pseudocode of duplicate-aware Micro- F_1 evaluation for ASTE.

```

num_pred = 0 # Count of predicted triplets
num_gold = 0 # Count of gold triplets
num_correct = 0 # Count of correct triplets

# Match predicted and gold triplets
# using set intersection
for sentence in data:
    pred_set = set(sentence.pred_triplets)
    gold_set = set(sentence.gold_triplets)
    correct_set = pred_set & gold_set

    num_pred += len(pred_set)
    num_gold += len(gold_set)
    num_correct += len(correct_set)

# Calculate scores
p = num_correct / num_pred # Precision
r = num_correct / num_gold # Recall
f1_score = 2 * p * r / (p + r)

```

A.2 Additional Hyperparameters

For GAS and Paraphrase models, there are 140M parameters when using BART-base. When using T5, there are 220M parameters. For BERT-base models (GTS, Span-ASTE, RoBMRC), there are roughly 110M parameters.

A.3 Effect of Sampling Size

For sampling number of sequence m , there are on average 3.12, 3.45, 3.84 unique triplets sampled for $m = 10, 20, 30$ respectively.

A.4 Annotation Guide

This section illustrates the guideline for human annotators. This task is a fine-grained sentiment analysis task where opinion terms, their aspect targets, and their expressed sentiments should be extracted

Name	Value
GPU Model	Nvidia A6000
CUDA Version	11.3
Python Version	3.7.12
PyTorch Version	1.11.0
ChatGPT API Cost	\$110
Generation Sampling Temperature	1.0

Table 7: List of experimental details.

together. Each sample contains one or multiple sentences which have been tokenized and labeled with indices. The annotation steps are as follows:

1. Read and understand the text sample and find out opinion terms as well as aspect target terms. Note that these terms should be explicit and the target term should not be a pronoun. If there is no opinion term or aspect target term, the sample is marked as “Invalid”.
2. If the sample contains opinion terms and aspect target terms, check whether there are aspect-opinion pairs. If not, the sample should also be marked as “Invalid”.
3. Determine the expressed sentiment of these pairs and record the spans of aspect-opinion pairs and their expressed sentiment in a 3-tuple format. Note that each sentence can have multiple triplets.

For example, given a review “The room was huge but terribly furnished”. We can find two aspect-opinion pairs (room, huge) with positive sentiment and (room, terribly furnished) with negative sentiment. The triplets of this text sample should be recorded in this format: ([1], [3], “POS”), ([1], [5, 6], “NEG”), where the index of the first token is 0.

There are several special cases that may make annotators hard to determine. We give a uniform guide here:

- Articles such as “the”, “a”, and “an” should not be included in target terms.
- Separate conjoined terms. For example, “The bedroom and washroom are big and clean”. “Bedroom and washroom” should be recorded as two separate terms “bedroom” and “washroom”. Opinion terms “big” and “clean” should also be separated.

Domain	Average Sample Length	POS%	NEU%	NEG%
Restaurant	16.37 tokens	73.01%	6.75%	20.24%
Laptop	18.36 tokens	57.50%	9.64%	32.86%
Hotel	21.92 tokens	59.25%	11.69%	29.06%
Cosmetics	21.61 tokens	45.68%	25.59%	28.74%

Table 8: More details of our domain-expanded ASTE dataset. We report the average length of samples and the percentage of positive (POS%), neutral (NEU%) and negative (NEG%) triplets respectively.

- It might be hard to determine whether some adverbs should be included in opinion terms. We should include these adverbs if they have a large influence on the sentiment polarity of the opinion term. For example, “This room is too big.” The opinion term should be “too big” instead of “big”, since “too” makes the opinion term express an obvious negative sentiment.

A.5 Detailed Results

A.6 Prompt Templates

To adapt ChatGPT to complex sentiment tasks such as ASTE, we design several templates based on previous works in generative ASTE (Zhang et al., 2021a).

A.7 More Details of Datasets

Table A.7 shows more details of our domain-expanded ASTE dataset. We can observe that our annotated hotel and cosmetics domains contain a larger average sample length and their label distribution is more balanced than previous restaurant and laptop domains.

A.8 Dataset Examples

Table 9 presents five examples for each domain. The standard of triplet formulation is the same across four domains and aspect target terms are domain-specific, indicating that our domain-expanded dataset can be well used as a cross-domain ASTE benchmark.

A.9 Case Study

Table 10 compares predictions of GAS and our GAS+CAGE method on two examples in two cross-domain settings. We find both methods show great performance in determining the sentiment. However, our method can identify the number of triplets more correctly, indicating that CAGE can effectively mitigate pseudo-label noise by reducing false positives and false negatives.

Domain	Example	Triplets
Restaurant	The service is awful .	(service, awful, negative)
	The chicken dinner was real good .	(chicken dinner, good, positive)
	The food is reliable and the price is moderate .	(food, reliable, positive), (price, moderate, neutral)
	Staffs are not that friendly , but the taste covers all .	(staffs, not that friendly, negative), (taste, covers all, positive)
Laptop	Prices are in line .	(prices, in line, neutral)
	The keyboard feels good and I type just fine on it .	(keyboard, good, positive)
	The battery gets so HOT it is scary .	(battery, HOT, negative), (battery, scary, negative)
	It 's great for streaming video and other entertainment uses .	(streaming video, great, positive), (entertainment uses, great, positive)
Hotel	This mouse is terrific .	(mouse, terrific, positive)
	Of course my warranty runs out next month .	(warranty, runs out, neutral)
	The smell was only slightly less prominent in our corner suite at the end of the hallway .	(smell, prominent, neutral)
	Also , the garbage trucks that frequent the ally are loud .	(garbage trucks, loud, negative)
Cosmetics	In the morning you can enjoy a free breakfast with many choices .	(breakfast, enjoy, positive), (breakfast, free, positive)
	The price was reasonable compared to the other options in the area .	(price, reasonable, positive)
	My fiancé opened the window shades and we had a huge brick wall for a view .	(brick wall, huge, neutral)
	It use to be one of the best products in the market .	(products, best, positive)
Cosmetics	This is a very heavy cover - up that feels heavy on your face .	(cover-up, heavy, neutral)
	Flimsy is really not a great thing when it 's 20 bucks .	(Flimsy, not a great thing, negative)
	I ordered the blonde color , but it really is a little dark .	(color, blonde, neutral), (color, dark, neutral)
	I love Essie but the formula on this one is awful .	(Essie, love, positive), (formula, awful, negative)

Table 9: Dataset examples.

	Hotel -> Cosmetics	Cosmetics -> Hotel
Example	Though it is more expensive than mass market gels , it does provide higher performance .	The rooms were very clean and the staff was very friendly and helpful especially when it came to ensuring we got on our buses for tours and our flights back home .
Gold label	(performance, higher, positive)	(rooms, clean, positive), (staff, friendly, positive), (staff, helpful, positive)
GAS prediction	(performance, higher, positive), (gels, expensive, negative)	(rooms, clean, positive), (staff, friendly, positive)
GAS+CAGE prediction	(performance, higher, positive)	(rooms, clean, positive), (staff, friendly, positive), (staff, helpful, positive)

Table 10: Case Study.