Self-Training with Pseudo-Label Scorer for Aspect Sentiment Quad Prediction

Yice Zhang[1,3], Jie Zeng[1*], Weiming Hu[1*], Ziyi Wang[1*],
Shiwei Chen[1,2], and Ruifeng Xu[1,2,3†]
[1] Harbin Institute of Technology, Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
zhangyc_hit@163.com, xuruifeng@hit.edu.cn

Abstract

Aspect Sentiment Quad Prediction (ASQP) aims to predict all quads (aspect term, aspect category, opinion term, sentiment polarity) for a given review, which is the most representative and challenging task in aspect-based sentiment analysis. A key challenge in the ASQP task is the scarcity of labeled data, which limits the performance of existing methods. To tackle this issue, we propose a self-training framework with a pseudo-label scorer, wherein a scorer assesses the match between reviews and their pseudo-labels, aiming to filter out mismatches and thereby enhance the effectiveness of self-training. We highlight two critical aspects to ensure the scorer's effectiveness and reliability: the quality of the training dataset and its model architecture. To this end, we create a human-annotated comparison dataset and train a generative model on it using ranking-based objectives. Extensive experiments on public ASQP datasets reveal that using our scorer can greatly and consistently improve the effectiveness of self-training. Moreover, we explore the possibility of replacing humans with large language models for comparison dataset annotation, and experiments demonstrate its feasibility.[1]

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) aims to recognize aspect-level opinions and sentiments from user-generated content (Pontiki et al., 2014). This problem has consistently attracted interest owing to its proficiency in distilling and summarizing fine-grained opinions from vast data (Do et al., 2019; Nazir et al., 2022; Zhang et al., 2023a). The most representative and challenging task in ABSA is Aspect Sentiment Quad Prediction (ASQP) (Cai et al., 2021; Zhang et al., 2021a). This task formulates aspect-level opinions and sentiments as quadruples, each consisting of an aspect term, aspect category, opinion term, and sentiment polarity. For example, given a review "*the food is great and reasonably priced*," the output of ASQP would be {(*food*, food_quality, *great*, positive), (*food*, food _prices, *reasonably priced*, positive)}.

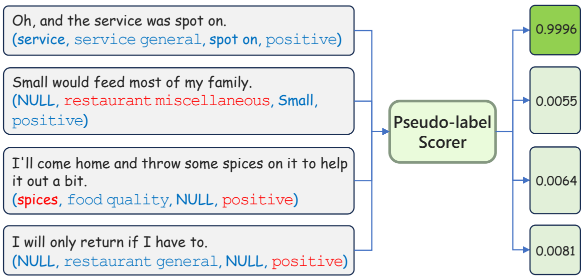

Oh, and the service was spot on.
(service, service general, spot on, positive)
Small would feed most of my family.
(NULL, restaurant miscellaneous, Small, positive)
I'll come home and throw some spices on it to help it out a bit.
(spices, food quality, NULL, positive)
I will only return if I have to.
(NULL, restaurant general, NULL, positive)
Pseudo-label Scorer
0.9996
0.0055
0.0064
0.0081

Figure 1: Illustration of our pseudo-label scorer.

As a fine-grained problem, ABSA faces the challenge of insufficient labeled data, which is particularly severe in the ASQP task. This issue limits the performance of existing models. Many efforts explore data augmentation methods to alleviate this issue. They synthesize new samples by modifying existing ones (Li et al., 2020; Hsu et al., 2021), applying self-training techniques (Wang et al., 2021), or utilizing generative methods (Yu et al., 2023; Deng et al., 2023; Zhang et al., 2023b). However, a significant limitation of these methods is that the synthetic samples often inevitably contain mismatches between sentences and labels, which can adversely affect model learning.

To reduce such mismatches, this paper introduces a pseudo-label scorer for data augmentation. As illustrated in Figure 1, the scorer assesses the degree of match between the review and its pseudo-label. If we have a sufficiently robust scorer, we can filter out all mismatched samples, thereby sig-

[*] The three authors contribute equally to this work.
[†] Corresponding Authors
[1] We release our code and data at https://github.com/HITSZ-HLT/ST-w-Scorer-ABSA.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11862–11875
August 11-16, 2024 ©2024 Association for Computational Linguistics

nificantly enhancing the effectiveness of data augmentation. We propose that the effectiveness and reliability of this scorer hinge on two critical aspects: (1) the quality of the training dataset and (2) its architecture along with the training objective. We discuss these two aspects below.

For the first aspect, previous works typically produce negative labels by modifying real labels using heuristic rules (Wang et al., 2021; Mao et al., 2022). However, such negative labels are usually simplistic and patterned, limiting the scorer's learning. To overcome this limitation, we create a human-annotated comparison dataset. Specifically, we train an ASQP model with existing labeled data, use it to infer several pseudo-labels for unlabeled data, and then have human annotators choose the most appropriate pseudo-labels. The labels chosen by annotators are designated as positive labels, while the rest as negative labels. Our dataset, in contrast to the rule-based datasets, is more challenging and better aligned with human judgment.

For the second aspect, previous works formalize label-scoring as a question-answering problem (Wang et al., 2021) or embed the discriminative matching token into the label (Mao et al., 2022). However, our findings suggest that these methods underperform in complex tasks like ASQP, due to their limited capacity to model the interactions between reviews and pseudo-labels. Recent works in preference optimization reveal that the language model itself can serve as a scorer (Rafailov et al., 2023; Yuan et al., 2023). This motivates us to use the conditional likelihoods that a generative model assigns to a pseudo-label as the measure of its quality. Compared with the previous methods, this approach enables the scorer to examine the plausibility of a pseudo-label in a token-by-token fashion, thus offering a more comprehensive and effective scoring. We then fine-tune this scorer on our comparison dataset using ranking-based objectives.

Upon developing this pseudo-label scorer, we apply it in a data augmentation framework, specifically opting for the self-training framework due to its simplicity. We conduct extensive experiments on public ASQP datasets to examine its effectiveness and further investigate the following questions: (1) how does the pseudo-label scorer perform using our comparison data and model architecture?; (2) is it feasible to replace humans with large language models to annotate the comparison data?; and (3) how to utilize the scorer to filter out low-quality samples? Furthermore, inspired by Ma et al. (2023), we extend the application of this scorer, employing it as a reranker for multiple candidate labels, and assess its impact and effectiveness.

Our contributions are summarized as follows:

(1) To the best of our knowledge, we are the first to apply a pseudo-label scorer to data augmentation in the ASQP task.

(2) We investigate how to enhance the scorer's effectiveness and reliability from both dataset and model architecture perspectives.

(3) We empirically demonstrate that the proposed pseudo-label scorer can significantly and consistently enhance the performance of existing models.

2 Background

Task Evolution. Aspect-Based Sentiment Analysis (ABSA) is a fine-grained sentiment analysis problem, aiming at recognizing user opinions and sentiments towards specific aspects (Zhang et al., 2023a). Within ABSA, aspects are generally defined as aspect categories or aspect terms. Aspect categories are predefined categories like food_quality and service_general in restaurant reviews, or laptop_portability and display_quality in laptop reviews. Aspect terms are explicit mentions of these aspect categories in the text. Based on this, Pontiki et al. (2014) formally define four subtasks: aspect category detection, aspect category sentiment classification, aspect term extraction, and aspect term sentiment classification. These tasks sequentially identify aspects and determine their corresponding sentiments.

Opinion term refers to words or phrases that express subjective sentiments. Sentiment expressions towards aspects often rely on these opinion terms, making them vital clues for recognizing aspects and their sentiment polarities. Consequently, many researchers focus on aspect and opinion co-extraction (Wang et al., 2016; Li and Lam, 2017; Fan et al., 2019; Chen et al., 2020; Zhao et al., 2020). Building on this, Peng et al. (2020) propose the Aspect Sentiment Triplet Extraction (ASTE) task, conceptualizing an aspect-level sentiment as a triplet consisting of an aspect term, opinion term, and sentiment polarity. Following ASTE, Cai et al. (2021); Zhang et al. (2021a) extend the triplet by incorporating the aspect category, evolving it into Aspect Sentiment Quad Prediction (ASQP). ASQP is currently the most representative and challenging task within ABSA.

ABSA Methods. Early efforts in ABSA primarily focus on identifying one or two sentiment elements. They mainly design specific model structures to establish interactions between sentiment elements and their contexts, as well as among different sentiment elements (Wang et al., 2016; Li and Lam, 2017; Ma et al., 2017; Chen et al., 2017; Xu et al., 2018; Xue and Li, 2018; Li et al., 2018; Fan et al., 2019; Zeng et al., 2019). Recent efforts have shifted towards compound tasks like ASTE and ASQP, proposing various end-to-end methods. These includes machine reading comprehension-based methods (Chen et al., 2021; Mao et al., 2021), table-filling methods (Wu et al., 2020; Chen et al., 2022; Zhang et al., 2022), span-based methods (Xu et al., 2021; Li et al., 2022; Cai et al., 2021), and generative methods (Yan et al., 2021; Zhang et al., 2021b,a; Mao et al., 2022; Gao et al., 2022; Hu et al., 2023a,b; Gou et al., 2023). Among these, generative methods have emerged as mainstream due to their universality and capacity to exploit rich label semantics.

Generative Aspect-Based Sentiment Analysis, abbreviated as GAS, is a unified generative framework proposed by Zhang et al. (2021b). The core idea of this framework is to transform sentiment elements into a label sequence and then use a SEQ2SEQ model to learn the dependencies between the input text and the label sequence. For the ASQP task, we can convert it into a label sequence using the following template:

$$\mathrm{seq}_{\mathrm{label}} = c_1 \mid s_1 \mid a_1 \mid o_1 \ ; \ \cdots \ ; \ c_n \mid s_n \mid a_n \mid o_n,$$

where c_i denotes the aspect category, s_i denotes the sentiment polarity, a_i denotes the aspect term, and o_i denotes the opinion term.

3 Comparison Dataset

We need to construct a comparison dataset to facilitate the training and evaluation of the pseudo-label scorer. This dataset comprises samples each containing a review sentence accompanied by several pseudo-labels, where one is the positive label and the others are negative labels. We train the scorer by requiring it to assign high scores to positive labels and low scores to negative labels.

Previous works typically produce negative labels using heuristic rules (Wang et al., 2021; Mao et al., 2022). They randomly modify elements in the existing positive labels, such as altering boundaries or conducting substitutions. Such negative labels are patterned and easily distinguishable, limiting the learning potential of the scorer. Therefore, this paper employs human annotators to construct this comparison dataset.

3.1 Data Preparation

Aligned with existing ASQP datasets (Cai et al., 2021; Zhang et al., 2021a), we collect reviews from two domains: `Restaurant` and `Laptop`. The restaurant reviews are from the Yelp Dataset[2], and the laptop reviews are from the Amazon Laptop Dataset[3] (Ni et al., 2019). We segment these reviews into individual sentences. Next, we employ the existing labeled dataset to train an ASQP model (Zhang et al., 2021b) and then utilize this model to generate four pseudo-labels for each review sentence via beam search.

3.2 Annotation Process

For a review sentence and its four pseudo-labels, annotators are presented with six options. The first four options correspond to the four pseudo-labels, the fifth option indicates that none of the pseudo-labels are appropriate, and the sixth option suggests that the review sentence does not express any sentiment or the expressed sentiment is difficult to infer. When annotators choose the fifth option, they are required to write an alternative label.

The annotation process is organized into multiple batches, each containing about 200 samples. To ensure accuracy, every sample is independently annotated by three different annotators. In cases of discrepancy among their annotations, a fourth annotator steps in to resolve the inconsistency. Furthermore, at the conclusion of each batch, the four annotators meet to discuss and reconcile any disagreements. Each annotator is provided with annotation guidelines[4] and existing labeled ASQP datasets. In instances where conflicts arise between the two, we prioritize adherence to the guidelines.

AI Annotation. Choosing the most appropriate label, while much simpler than annotating an ASQP

[2] https://www.yelp.com/dataset

[3] https://nijianmo.github.io/amazon/index.html

[4] https://alt.qcri.org/semeval2014/task4/data/uploads/semeval14_absa_annotationguidelines.pdf and https://alt.qcri.org/semeval2016/task5/data/uploads/absa2016_annotationguidelines.pdf.

Datasets	P1	P2	P3	P4	P5	P6	Total
ACOS-Laptop-Comp	853	149	88	44	167	204	1505
ACOS-Rest-Comp	894	109	49	23	122	1003	2200
ACOS-Laptop-Comp-AI	1744	410	213	122	0	259	2748
ACOS-Rest-Comp-AI	1796	276	110	60	0	1620	3862
ASQP-Rest15-Comp-AI	1585	369	157	94	0	1223	3428
ASQP-Rest16-Comp-AI	1560	431	114	76	0	1337	3518

Table 1: Statistics of the comparison datasets. P1-P6 correspond to the number of samples for options 1 to 6, respectively.

label from scratch, remains a laborious task. Therefore, we explore the feasibility of using ChatGPT as a substitute for human annotators. To ensure the quality of AI annotations, we carefully craft prompts for each ASQP dataset. Additionally, we incorporate three strategies to enhance the annotation process: self-consistency, self-assessment, and rationale augmentation. Details of AI annotation can be found in Appendix A.

3.3 Statistics

We construct two human-annotated and four AI-annotated comparison datasets. Their basic statistical information is presented in Table 1. In the training phase of the scorer, we exclude samples corresponding to option 6 and reserve a portion of the data as the development set for hyperparameter tuning and model selection. Specifically, for the Restaurant datasets, we set aside 200 samples, and for the Laptop datasets, we allocate 300 samples. Consequently, this leaves around 1,000 training samples in the human-annotated datasets and approximately 2,000 training samples in the AI-annotated datasets.

4 Our Approach

4.1 Pseudo-label Scorer

The objective of the pseudo-label scorer is to score the match between a review and a pseudo-label. Previous efforts formalize this scoring task as a question-answering problem (Wang et al., 2021) or embed the discriminative matching token into the label (Mao et al., 2022). However, these methods struggle to effectively capture the interaction between reviews and pseudo-labels. Inspired by recent works in preference optimization (Rafailov et al., 2023; Yuan et al., 2023; Song et al., 2023), we utilize a generative model as the scorer. Given a review sentence x and a pseudo label y, their matching score is quantified by the conditional probability assigned by the generative model:

$$s(x, y) \propto p(y|x) = \prod_t p(y_t|y_{<t}, x). \quad (1)$$

Compared to previous methods, this approach integrates the likelihood of each token in the pseudo-label to derive its overall score, thereby providing a comprehensive and effective scoring.

Training. We optimize the pseudo-label scorer on the annotated comparison dataset with the ranking-based training objective. Specifically, we design a simple listwise objective[5] as follows:

$$\mathcal{L}_{\text{LIST}} = -\log \frac{p(y_p|x)}{Z}, \quad (2)$$

$$Z = p(y_p|x) + \sum_{y_n} p(y_n|x), \quad (3)$$

where y_p denotes the positive label, y_n denotes the negative label, and Z is the normalization factor.

In addition to the comparison dataset, we also incorporate the original ASQP dataset to further enhance the training of the scorer. Labels from the original ASQP dataset are treated as additional positive labels and are combined with the positive labels from the comparison dataset. We additionally maximize the scores of these positive labels to enhance the scorer. The combined loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_1 + \alpha \mathcal{L}_2, \quad (4)$$

$$\mathcal{L}_1 = \mathbb{E}_{(x,\mathcal{Y}) \sim D_{\text{COMP}}} \mathcal{L}_{\text{LIST}}(x, \mathcal{Y}), \quad (5)$$

$$\mathcal{L}_2 = \mathbb{E}_{(x,y_p) \sim D_{\text{COMP}} \cup D_{\text{ASQP}}} - \log p(y_p|x), \quad (6)$$

where D_{COMP} represents the comparison dataset, D_{ASQP} represents the original ASQP dataset, $\mathcal{Y}$ denotes the set of several pseudo-labels of the sentence x, and α is a hyperparameter.

4.2 Self-Training with Data Filtering

Self-training (Scudder, 1965), a simple and classic semi-supervised technique, can be applied for data augmentation. It consists of three main steps: (1) training an initial model with the existing labeled dataset, (2) using this model to generate pseudo-labels for unlabeled data, and (3) finally incorporating these pseudo-labeled data into the labeled dataset. However, this method inevitably introduces low-quality pseudo-labels, where the label does not accurately match the given review. To

[5] Besides the listwise objective, we also explore pointwise and pairwise objectives, which are presented in Appendix B.

overcome this issue, we implement a two-stage filtering process that leverages both the initial model and the pseudo-label scorer.

Confidence-based Filtering. We first use the confidence of the initial model in the pseudo-label as the measure of its quality. Thus, we filter out those samples with minimum confidence below a certain threshold. Formally, we retain samples (x, y) satisfying

$$\left[\min_t p(y_t|y_{<t}, x)\right] \geq \gamma_1, \tag{7}$$

where γ_1 is a hyper-parameter and is empirically set to 0.7.

Scorer-based Filtering. Next, we use the pseudo-label scorer to evaluate the remaining samples. We observe that pseudo-labels with low scores are consistently of poor quality. Besides, while samples with high scores generally exhibit good label quality, their sentences tend to be overly simple, offering limited helpfulness for subsequent model training. Therefore, we retain only those samples whose scores fall between thresholds γ_2 and γ_3, which can be formulated as follows:

$$\gamma_2 \leq s(x, y) \leq \gamma_3. \tag{8}$$

4.3 Pseudo-label Scorer as Reranker

Reranking is originally a concept in information retrieval, referring to the process of rescoring and reranking preliminary candidate results. Ma et al. (2023) show that incorporating a reranking step can enhance performance in information extraction tasks. In this paper, we claim that our pseudo-label scorer can serve as such a reranker. Specifically, for a given review, we first utilize an ASQP model to generate four candidate labels via beam search and then select the best one from these candidates using our pseudo-label scorer. The selected candidate is utilized as the final output.

5 Experiments

5.1 Experiment Setup

Datasets. We evaluate our approach on four public ASQP datasets. These datasets originate from the SemEval Challenges (Pontiki et al., 2015, 2016) and Amazon platform during 2017 and 2018. The quad-level annotations are provided by Cai et al. (2021) and Zhang et al. (2021a). Detailed statistics of these datasets are presented in Table 2. Besides, to train the pseudo-label scorer, we construct several comparison datasets, the statistics of which can be found in Table 1.

Datasets	Train		Dev		Test	
	#S	#Q	#S	#Q	#S	#Q
`ACOS-Laptop`	2934	4172	326	440	816	1161
`ACOS-Rest`	1530	2484	171	261	583	916
`ASQP-Rest15`	834	1354	209	347	537	795
`ASQP-Rest16`	1264	1989	316	507	544	799

Table 2: Statistics of four ASQP datasets (Cai et al., 2021; Zhang et al., 2021a). #S and #Q represent the number of sentences and quads.

Implementation Details. We utilize `T5-large` (Raffel et al., 2020) as the backbone for our pseudo-label scorer. During the training phase, we set both the batch size and the number of training epochs to 10. For other hyperparameters, including the learning rate and α, we perform a simple hyperparameter search. Once the scorer is trained, we apply it to score and rank the pseudo-labeled samples[6]. For datasets `ACOS-Rest`, `ASQP-Rest15`, and `ASQP-Rest16`, we retain samples with scores falling within the top 10% to 40%; for the `ACOS-Laptop` dataset, this range is set from 20% to 50%. From these retained samples, we randomly select 10,000 samples and merge them with the original labeled dataset to form the augmented dataset. To reduce the impact of randomness, we run our approach five times and report the average results.

Baselines. To validate the effectiveness of the proposed approach, we integrate it into two typical ASQP methods: GAS (Zhang et al., 2021b) and MUL (Hu et al., 2023b). We run these two methods on the augmented dataset and incorporate a reranking step during the inference phase to enhance the predictions. We also benchmark our approach against a range of other methods, including Extract-Classify (Cai et al., 2021), Paraphrase (Zhang et al., 2021a), Seq2Path (Mao et al., 2022), DLO/ILO (Hu et al., 2022), LEGO-ABSA (Gao et al., 2022), MvP (Gou et al., 2023), GenDA (Wang et al., 2023), and ChatGPT (few-shot) (Xu et al., 2023).

[6] The source for the pseudo-labeled data is identical to that for the comparison data, originating from the Yelp Dataset and Amazon Laptop Dataset. But there is no overlap between them. The ASQP model used for pseudo-labeling is GAS (Zhang et al., 2021b).

Objectives	ACOS-Laptop	ACOS-Rest
Wang et al. (2021)	50.67	56.10
Mao et al. (2022)	64.34	72.00
Ours	**67.74**	**78.50**

Table 3: Comparison results of the architecture for the pseudo-label scorer (accuracy, %). Wang et al. (2021) formalize label-scoring as a question-answering problem. Mao et al. (2022) append a discriminative matching token to the label.

Annotation Schemes	ACOS-Laptop	ACOS-Rest
NONE	60.53	74.10
HUMANN-1234	63.60	75.00
HUMANN-12345	64.60	76.60
HUMANN-12345*	67.74	78.50
AIANN-1234*	**68.67**	**79.60**

Table 4: Experimental results of annotating the comparison dataset (accuracy, %): (1) NONE denotes the approach where neither human nor AI annotations are used, and the pseudo-label with the highest model confidence is selected as the positive label; (2) HUMANN-1234 represents the annotation scheme where human annotators choose the best pseudo-label out of four; (3) HUMANN-12345 extends HUMANN-1234 by allowing human annotators to write an additional label when none of the four options are suitable; (4) AIANN-1234 mirrors HUMANN-1234, but with ChatGPT replacing human annotators; (5) methods with * indicate the training of the scorer using both the comparison dataset and the original ASQP dataset.

Datasets	w/ P6		w/o P6	
	Kappa	Accu	Kappa	Accu
ACOS-Laptop-Comp-AI	62.71	79.20	65.84	86.30
ACOS-Rest-Comp-AI	67.44	80.90	47.12	87.15

Table 5: Consistency between AI- and human-annotated comparison Data (%). P6 refers to samples with option 6 selected. We calculate the consistency both before and after removing these samples.

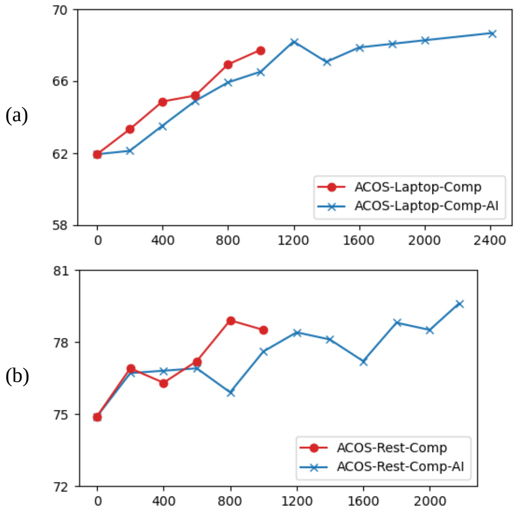


Figure 2: Performance trends of comparison data with increasing data quantity (accuracy, %): (a) results on ACOS-Laptop; (b) results on ACOS-Rest.

5.2 Analysis of Pesudo-label Scorer

Given the importance of the pseudo-label scorer in our framework, we first undertake an analysis of it, focusing on two key aspects: its model architecture and the training dataset.

Model Architecture. We use the conditional likelihood the generative model assigns to a pseudo-label as its scoring metric. To examine the effectiveness of our approach, we conduct experiments on two human-annotated comparison datasets and benchmark our approach against previous methods (Wang et al., 2021; Mao et al., 2022). As Table 3 illustrates, previous methods, especially the question-answering method, perform poorly in the ASQP task. In contrast, our approach achieves a significant advantage, demonstrating its effectiveness.

Comparison Dataset. We conduct experiments to compare different annotation schemes and list the results in Table 4. We have the following observations. (1) Utilizing humans or AI to annotate the comparison data is crucial, as their performance is noticeably superior to that without annotation. Particularly, allowing human annotators to write a label when no option is suitable can notably enhance performance. (2) Training the scorer with a combination of the comparison data and the original ASQP dataset is more effective than using the comparison data alone. (3) AI-annotated comparison data can achieve even better results than human-annotated comparison data.

We conduct a further analysis of AI Annotation. Table 5 presents the consistency between AI- and human-annotated data. Although the consistency is not very high statistically, considering the subjective nature of this task, the quality of AI annotation is acceptable. Additionally, a significant advantage of AI annotation lies in its cost-effectiveness relative to human annotation, enabling the efficient acquisition of a large amount of annotated data.

Figure 2 illustrates the performance trends of human- and AI-annotated data relative to their quantities. Although AI-annotated data exhibits

Methods	ACOS-Laptop			ACOS-Rest			ASQP-Rest15			ASQP-Rest16		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Extract-Classify (Cai et al., 2021)	45.56	29.28	35.80	38.54	52.96	44.61	35.64	37.25	36.42	38.40	50.93	43.77
Paraphrase (Zhang et al., 2021a)	-	-	-	-	-	-	46.16	47.72	46.93	56.63	59.30	57.93
Seq2Path (Mao et al., 2022)	-	-	42.97	-	-	58.41	-	-	-	-	-	-
DLO (Hu et al., 2022)	43.40	43.80	43.60	60.02	59.84	59.18	47.08	49.33	48.18	57.92	61.80	59.79
ILO (Hu et al., 2022)	44.14	44.56	44.35	58.43	58.95	58.69	47.78	50.38	49.05	57.58	61.17	59.32
LEGO-ABSA (Gao et al., 2022)	-	-	-	-	-	-	-	-	45.80	-	-	57.70
MvP (Gou et al., 2023)	-	-	43.92	-	-	61.54	-	-	51.04	-	-	60.39
GenDA (Wang et al., 2023)	-	-	-	-	-	-	49.74	50.29	50.01	60.08	61.70	60.88
ChatGPT (few-shot) (Xu et al., 2023)	21.72	27.65	24.33	38.39	46.40	42.02	29.66	37.86	33.26	36.09	46.93	40.81
GAS (Zhang et al., 2021b)	43.46	42.69	43.07	59.81	57.51	58.63	47.15	46.01	46.57	57.30	57.82	57.55
+ ST	44.35	42.75	43.54	59.95	58.98	59.46	46.86	47.65	47.25	57.95	58.75	58.35
+ ST & C-Filter	45.14	44.00	44.56	60.57	58.82	59.67	49.04	47.77	48.40	59.18	59.72	59.45
+ ST & CS-Filter	46.23	44.41	45.30	63.41	60.00	61.66	-	-	-	-	-	-
+ ST & CS-Filter & ReRank	**46.76**	**45.00**	**45.86**	**64.66**	**61.33**	**62.95**	-	-	-	-	-	-
+ ST & CS-Filter (AI)	46.44	44.01	45.19	62.69	60.24	61.44	50.92	49.86	50.38	60.87	61.30	61.08
+ ST & CS-Filter & ReRank (AI)	**47.00**	**45.05**	**46.01**	**63.74**	**61.25**	**62.47**	**51.59**	**51.90**	**51.74**	**62.55**	**64.31**	**63.51**
MUL (Hu et al., 2023b)	44.38	43.65	44.01	61.22	59.87	60.53	49.12	50.39	49.75	59.24	61.75	60.47
MUL (*Our Reproduction*)	42.79	41.95	42.45	61.22	59.80	60.50	48.28	49.74	48.99	58.42	60.68	59.52
+ ST	43.38	42.98	43.23	61.17	59.89	60.67	47.94	49.21	48.57	57.45	59.37	58.39
+ ST & C-Filter	44.59	43.67	44.13	62.11	60.20	61.14	48.74	49.06	48.90	59.44	61.07	60.25
+ ST & CS-Filter	44.67	43.72	44.19	63.57	60.67	62.09	-	-	-	-	-	-
+ ST & CS-Filter & ReRank	**46.88**	**44.74**	**45.78**	**66.18**	**61.75**	**63.89**	-	-	-	-	-	-
+ ST & CS-Filter (AI)	44.89	44.07	44.47	64.28	61.31	62.76	50.78	51.17	50.97	61.39	62.68	62.03
+ ST & CS-Filter & ReRank (AI)	**47.05**	**45.32**	**46.17**	**65.43**	**61.92**	**63.63**	**51.94**	**52.00**	**51.97**	**63.46**	**64.31**	**63.88**

Table 6: Experimental results on four ASQP datasets (%). C-Filter indicates the application of confidence-based filtering. CS-Filter represents the integration of confidence-based and scorer-based filtering. Methods marked with AI indicate that the pseudo-label scorer used is trained on AI-annotated comparison data.

lower performance at the same quantity, the scalability of AI annotation allows it to catch up and potentially exceed human-annotated data's performance when using more data. For instance, more than 2,000 AI-annotated samples can equal or outperform 1,000 human-annotated samples. Consequently, we can conclude that for the ASQP task, it is feasible to replace humans with AI to annotate the comparison data.

5.3 Analysis of Self-Training

Main Results. We develop a self-training framework using the pseudo-label scorer, with the experimental results presented in Table 6. According to these results, our approach substantially and consistently improves the performance of existing ASQP methods (Zhang et al., 2021b; Hu et al., 2023b). Specifically, GAS achieves F_1-score improvements of 2.94%, 4.32%, 5.17%, and 5.96% across the four datasets, averaging at 4.60%; MUL achieves F_1-score improvements of 3.72%, 3.39%, 2.98%, and 4.36% across these datasets, averaging at 3.61%. Upon integrating our approach, both GAS and MUL outperform previous methods. These results demonstrate the effectiveness of our approach.

Furthermore, we have the following observations. (1) The two-stage filtering process, namely CS-Filter, greatly enhances the effectiveness of self-training. In most datasets, it results in over 2% improvement compared to self-training alone, highlighting the importance of data filtering in the self-training framework. (2) Incorporating a rerank step can further improve performance, by around 1%. (3) Using AI-annotated data in downstream self-training can attain results comparable to those using human-annotated data. This further indicates the feasibility of replacing human annotators with AI for comparison data annotation. (4) ChatGPT performs poorly on the ASQP task, suggesting that using it directly for this task does not fully leverage its capabilities. Conversely, using it for comparison data annotation effectively exploits its strengths. (5) It can be noted that our filtering strategy offers relatively limited improvements on `ACOS-Laptop`. We attribute this to potential inconsistency between its ASQP annotations and our comparison annotations. A more detailed discussion is available in Further Analysis.

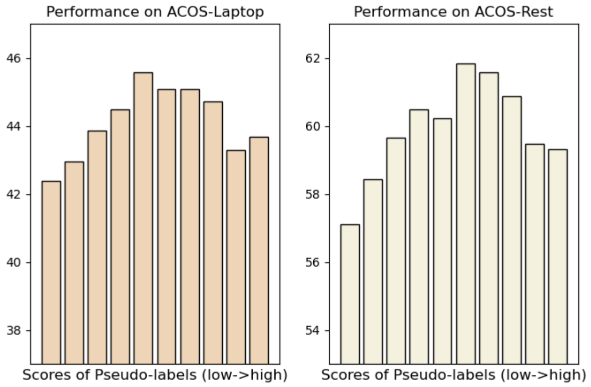


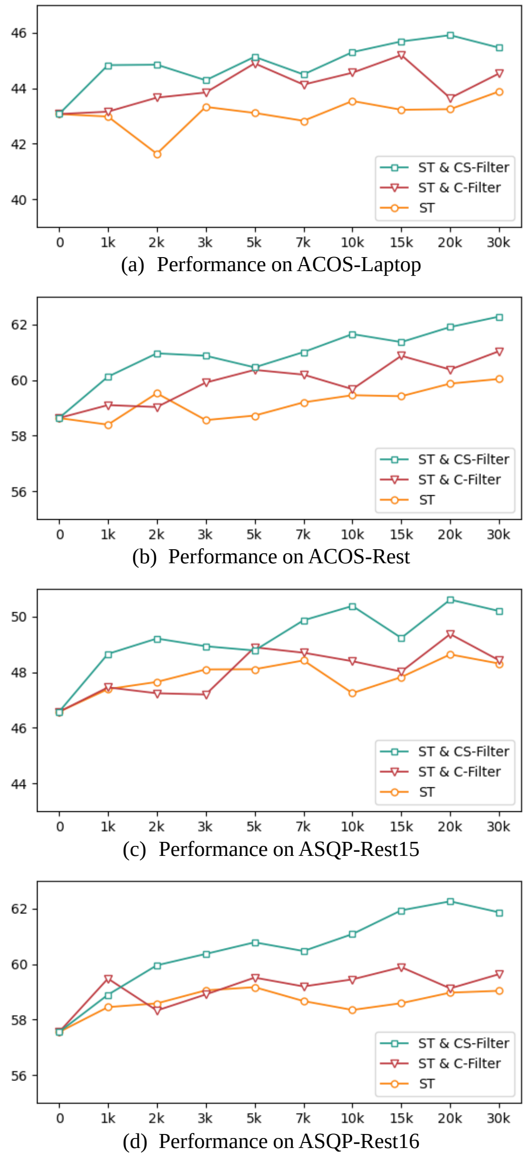
Figure 3: Performance of GAS on the augmented dataset under different match scores (F_1-score, %).

Effect of Match Score. Our approach relies on the match scores output by the pseudo-label scorer for data filtering. We conduct experiments to examine the impact of these scores on self-training performance. Figure 3 illustrates that the performance incrementally increases with increasing match scores. Nevertheless, beyond a certain threshold, further increases in match scores lead to a decline in performance. This phenomenon confirms our hypothesis that samples with too low scores suffer from poor label quality, adversely affecting model learning, and samples with too high scores tend to be overly simple, providing limited helpfulness for subsequent model training.

Effect of Data Quantity. The quantity of pseudo-labeled samples is another important factor in the effectiveness of self-training. We conduct experiments to analyze its impact. As illustrated in Figure 4, there is an overall upward trend in performance with increased data quantity. Notably, this trend is more stable and pronounced following two-stage filtering, underscoring the necessity of data filtering. Furthermore, we notice a decrease in self-training performance when the number of augmented samples exceeds 20,000. This suggests that there is a limit to improving performance by simply increasing data quantity. Balancing diversity and label quality to enhance the effectiveness of self-training warrants further exploration in subsequent research.

5.4 Further Analysis

Comparison Data as Additional Labeled Data. One feasible approach for utilizing comparison data is to treat each sample along with its positive label as an additional labeled ASQP sample. We analyze the effectiveness of this approach, and



(a) Performance on ACOS-Laptop

(b) Performance on ACOS-Rest

(c) Performance on ASQP-Rest15

(d) Performance on ASQP-Rest16

Figure 4: Performance of GAS under different numbers of augmented samples (F_1-score, %).

the results in Table 7 reveal: (1) this approach can improve performance, with human-annotated comparison data outperforming AI-annotated data; (2) under conditions of equal data volume, it is superior to self-training without data filtering, indicating that the quality of comparison data is better than that of pseudo-labeled data; and (3) however, it falls significantly short of self-training with data filtering. These findings suggest that utilizing comparison data to train a pseudo-label scorer is more effective than simply treating it as additional labeled data.

Pseudo-label Scorer as the ASQP Model. The pseudo-label scorer is architecturally a generative model and can potentially be used as an ASQP

Methods	Laptop	Rest	Rest15	Rest16	Avg
GAS	43.07	58.63	46.57	57.55	-
+ ST & C-Filter (1k)	43.16	59.10	47.45	59.48	+0.84
+ ST & C-Filter (2k)	43.66	59.03	47.24	58.33	+0.61
+ ST & C-Filter (10k)	44.56	59.67	48.40	59.45	+1.57
+ ST & CS-Filter (10k)	45.30	61.66	50.38	61.08	+3.15
+ CompData (1k)	43.51	60.20	-	-	-
+ CompData (2k, AI)	42.99	59.77	48.15	59.73	+1.20

Table 7: Experimental results of using comparison data as additional labeled data (F_1-scorer, %). The human- and AI-annotated comparison datasets contain about 1,000 and 2,000 samples, respectively.

Methods	Laptop	Rest	Rest15	Rest16
GAS	43.07	58.63	46.57	57.55
+ Our Approach	45.86	62.95	-	-
+ Our Approach (AI)	46.01	62.47	51.74	63.51
Pseudo-Label Scorer	43.93	60.66	-	-
Pseudo-Label Scorer (AI)	44.06	60.00	51.59	61.49

Table 8: Experimental results of using pseudo-label scorer as the ASQP model (F_1-scorer, %).

model. We evaluate this possibility and list the results in Table 8. A surprising finding is that directly using the scorer to predict quads achieves good performance, though it generally falls short of using it for filtering and ranking. This suggests that, besides training the scorer, leveraging comparison data to enhance the ASQP model is a promising direction, deserving of in-depth investigation in future research.

Assessing Label Quality in ASQP Data. Beyond assessing pseudo-labeled data, our scorer can assess the quality of existing ASQP data. We conduct a statistical analysis of match scores for ASQP samples and list the results in Table 9. Our analysis reveals relatively low match scores for the `ACOS-laptop` dataset, suggesting either poor annotation quality or low consistency with our comparison data. We manually review 100 samples with match scores low 0.1 and find that 73% of the data contradicts the annotation guidelines, including 44% with errors in aspect category annotations, 8% in aspect or opinion term annotations, and 6% in sentiment annotations. Moreover, we experiment with removing samples with low match scores. The results presented in Table 10 show that this removal not only preserves model performance but enhances it. These findings indicate that our scorer is an effective tool for assessing label quality in existing datasets and that the removal of low-quality samples is advantageous.

	Laptop	Rest	Rest15	Rest16
< 0.1	19.12%	3.92%	3.24%	1.50%
< 0.3	30.30%	7.32%	5.52%	2.37%
< 0.5	40.52%	10.46%	7.07%	3.56%
< 0.7	52.56%	14.58%	9.59%	5.46%
< 0.9	70.25%	25.88%	15.83%	9.43%

Table 9: Statistics of match scores in the ASQP training datasets.

Ratio of Removal	Laptop	Rest	Rest15	Rest16	Avg
0%	43.07	58.63	46.57	57.55	-
2%	43.38	59.44	47.20	58.27	+0.62
4%	43.50	**59.51**	**48.09**	59.10	**+1.10**
6%	43.89	59.24	46.78	58.25	+0.59
8%	43.42	59.38	46.68	58.43	+0.52
10%	**44.25**	59.09	46.57	**59.20**	+0.82

Table 10: Performance after removing samples with low match scores in the training set (F_1-scorer, %).

Analysis of Reranking. We present the analysis of the reranking step in Appendix C.

6 Conclusions

In this paper, we introduce a pseudo-label scorer for the Aspect Sentiment Quad Prediction (ASQP) task to reduce mismatches in data augmentation. We propose that the effectiveness and reliability of this scorer hinge on two critical aspects: the quality of the training dataset and its model architecture. To this end, we create both human- and AI-annotated comparison datasets and propose a scoring method based on a generative model. Upon developing this scorer, we apply it to data filtering in a self-training framework and further employ it as a reranker to enhance ASQP models. Detailed experiments and analysis demonstrate the effectiveness of our comparison datasets and the proposed architecture. Furthermore, experimental results on four public ASQP datasets reveal that our scorer significantly and consistently improves the performance of existing methods.

Acknowledgements

We thank the anonymous reviewers for their valuable suggestions to improve the quality of this work. We also thank Qianlong Wang for the enlightening discussions that significantly contributed to this work. This work was partially supported by the National Natural Science Foundation of China 62176076, Natural Science Foundation of Guangdong 2023A1515012922, the Shenzhen Foundational Research Funding JCYJ20220818102415032, the Major Key Project of PCL2021A06, and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

Limitations

While our approach significantly enhances the effectiveness of data augmentation and improves the performance of existing ASQP models, it also suffers from the following limitations:

- Data augmentation generally comprises two pivotal components: data synthesis and quality control. While this paper focuses primarily on the latter, the former is equally vital for the success of data augmentation. Given that models trained on limited labeled data may underperform in certain categories or contexts, targeted data synthesis can mitigate these issues. A comprehensive exploration of both data synthesis and quality control is essential for developing an effective and robust data augmentation framework.
- The implementation of our approach necessitates manually annotated comparison data. Although we could use large language models to replace human annotators, crafting and refining prompts still demands meticulous human expertise and is notably time-intensive.

We argue that these limitations offer promising directions for future research.

References

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.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985, Dublin, Ireland. Association for Computational Linguistics.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 452–461, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12666–12674.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Bidirectional generative framework for cross-domain aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12272–12285, Toronto, Canada. Association for Computational Linguistics.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications*, 118:272–299.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. 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 2509–2518, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7002–7012, Gyeongju, Republic of Korea. International Committee on 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.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Semantics-preserved data augmentation for aspect-based sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4417–4422, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hanxu Hu, Yunqing Liu, Zhongyi Yu, and Laura Perez-Beltrachini. 2023a. Improving user controlled table-to-text generation robustness. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2317–2324, Dubrovnik, Croatia. Association for Computational Linguistics.

Mengting Hu, Yinhao Bai, Yike Wu, Zhen Zhang, Liqi Zhang, Hang Gao, Shiwan Zhao, and Minlie Huang. 2023b. 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.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7056–7066, Online. Association for Computational Linguistics.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956, Melbourne, Australia. Association for Computational Linguistics.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892, Copenhagen, Denmark. Association for Computational Linguistics.

You Li, Yongdong Lin, Yuming Lin, Liang Chang, and Huibing Zhang. 2022. A span-sharing joint extraction framework for harvesting aspect sentiment triplets. *Knowledge-Based Systems*, 242:108366.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4068–4074. AAAI Press.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples!

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.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13543–13551.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2022. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, Hong Kong, China. Association for Computational Linguistics.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International*

Workshop on Semantic Evaluation (SemEval-2016), pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

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. *J. Mach. Learn. Res.*, 21(1).

H. Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment.

An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, and Naoaki Okazaki. 2023. Generative data augmentation for aspect sentiment quad prediction. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 128–140, Toronto, Canada. Association for Computational Linguistics.

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.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626, Austin, Texas. Association for Computational Linguistics.

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.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Melbourne, Australia. 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.

Xiancai Xu, Jia-Dong Zhang, Rongchang Xiao, and Lei Xiong. 2023. The limits of chatgpt in extracting aspect-category-opinion-sentiment quadruples: A comparative analysis.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia. 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.

Jianfei Yu, Qiankun Zhao, and Rui Xia. 2023. Cross-domain data augmentation with domain-adaptive language modeling for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1456–1470, Toronto, Canada. Association for Computational Linguistics.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16).

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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023a. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

Yice Zhang, Yifan Yang, Meng Li, Bin Liang, Shiwei Chen, and Ruifeng Xu. 2023b. Target-to-source augmentation for aspect sentiment triplet extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12165–12177, Singapore. Association for Computational Linguistics.

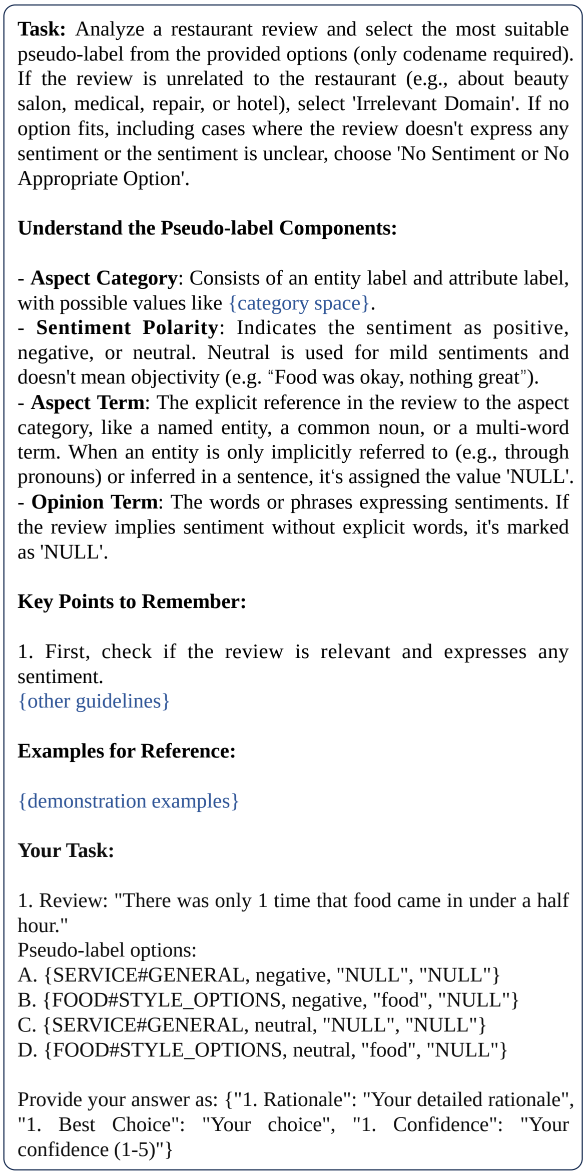
Yice Zhang, Yifan Yang, Yihui Li, Bin Liang, Shiwei Chen, Yixue Dang, Min Yang, and Ruifeng Xu. 2022. Boundary-driven table-filling for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6485–6498, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248, Online. Association for Computational Linguistics.

A Details of AI Annotation

We employ ChatGPT[7] to replace humans for comparison data annotation. The prompt is depicted in Figure 5. We continuously refine the guidelines and demonstrations in the prompt based on the annotation results. Furthermore, to further enhance annotation quality, we adopt three strategies:

- **Self-consistency**: For each sample, we input it into ChatGPT twice. We keep the sample if the results are consistent.
- **Self-assessment**: We require ChatGPT to evaluate its confidence level on a scale from 1 to 5 after each judgment. We retain only those samples with a confidence level is 5.
- **Rationale augmentation**: We instruct ChatGPT to provide reasoning and explanation before making its judgment.



Task: Analyze a restaurant review and select the most suitable pseudo-label from the provided options (only codename required). If the review is unrelated to the restaurant (e.g., about beauty salon, medical, repair, or hotel), select 'Irrelevant Domain'. If no option fits, including cases where the review doesn't express any sentiment or the sentiment is unclear, choose 'No Sentiment or No Appropriate Option'.

Understand the Pseudo-label Components:

- **Aspect Category**: Consists of an entity label and attribute label, with possible values like {category space}.
- **Sentiment Polarity**: Indicates the sentiment as positive, negative, or neutral. Neutral is used for mild sentiments and doesn't mean objectivity (e.g. "Food was okay, nothing great").
- **Aspect Term**: The explicit reference in the review to the aspect category, like a named entity, a common noun, or a multi-word term. When an entity is only implicitly referred to (e.g., through pronouns) or inferred in a sentence, it's assigned the value 'NULL'.
- **Opinion Term**: The words or phrases expressing sentiments. If the review implies sentiment without explicit words, it's marked as 'NULL'.

Key Points to Remember:

1. First, check if the review is relevant and expresses any sentiment.
{other guidelines}

Examples for Reference:

{demonstration examples}

Your Task:

1. Review: "There was only 1 time that food came in under a half hour."
Pseudo-label options:
A. {SERVICE#GENERAL, negative, "NULL", "NULL"}
B. {FOOD#STYLE_OPTIONS, negative, "food", "NULL"}
C. {SERVICE#GENERAL, neutral, "NULL", "NULL"}
D. {FOOD#STYLE_OPTIONS, neutral, "food", "NULL"}

Provide your answer as: {"1. Rationale": "Your detailed rationale", "1. Best Choice": "Your choice", "1. Confidence": "Your confidence (1-5)"}

Figure 5: Prompt for AI Annotation in `ACOS-Rest`.

Additionally, to reduce annotation costs, we integrate four samples into one prompt and annotate them at once.

B Ranking Objectives for Training Scorer

We optimize the pseudo-label scorer on the annotated comparison dataset and explore three ranking-based training objectives: pointwise, pairwise, and listwise approaches. The pointwise approach classifies positive and negative samples separately and

[7] Available at `https://chat.openai.com/`. The specific model used is `gpt-4-1106-preview`.

can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{POINT}} = &-\log p(y_p|x) \\ &-\sum_{y_n} \log(1-p(y_n|x)),\end{aligned} \tag{9}$$

where y_p denotes the positive label, and y_n denotes the negative label. The pairwise approach focuses on the relative quality of labels. We implement two pairwise training objectives, detailed as follows:

$$\mathcal{L}_{\text{PAIR1}} = -\sum_{y_n} \log \sigma \left[\beta \log \frac{p(y_p|x)}{p(y_n|x)}\right], \tag{10}$$

$$\mathcal{L}_{\text{PAIR2}} = \sum_{y_n} \max(0, p(y_n|x) - p(y_p|x)), \tag{11}$$

where β is a hyper-parameter. Lastly, the listwise approach optimizes the ranking of the entire list. We design a simple listwise training objective, as outlined in Equation 2.

B.1 Experiment Results

Objectives	ACOS-Laptop	ACOS-Rest
Pointwise	**67.93**	77.80
Pairwise1	67.13	77.00
Pairwise2	66.87	77.40
Listwise	67.74	**78.50**

Table 11: Comparison results of four ranking-based objectives (accuracy, %).

Experimental results in Table 11 reveal that pointwise and listwise objectives outperform two pairwise objectives, with the listwise objective being slightly better overall. Consequently, we adopt the listwise objective as the default training objective.

C Analysis of Reranking

	Laptop	Rest	Rest15	Rest16	Avg
ST-CS	45.30	61.66	50.38	61.08	54.61
ST-CS & RERANK	45.86	62.95	51.74	63.51	+1.41
ST-CS & RERANK♮	63.22	75.63	66.27	76.29	+15.75

Table 12: Experimental results of reranking (F_1-score, %). ST-CS denotes self-training with confidence- and scorer-based filtering. ♮ indicates the performance achieved using a perfect reranker.

We apply our pseudo-label scorer as a reranker to rescore the candidate labels generated by the ASQP model, selecting the highest-scoring one as the final result. Table 12 shows that this reranking step significantly improves performance on the ASQP task, resulting in an average F_1 improvement of 1.41%. Additionally, if we consider a hypothetical perfect reranker that always selects the optimal candidate, Table 12 shows that the performance gain could reach up to 15.75%. This significant potential boost underscores the value of further exploring the reranking step.

	Laptop	Rest	Rest15	Rest16
	Best Candidates			
1-st	67.70%	69.37%	61.15%	65.11%
2-nd	13.77%	15.44%	17.17%	16.29%
3-rd	10.44%	8.64%	13.48%	11.14%
4-th	8.09%	6.55%	8.19%	7.46%
	Scorer's Preferred Choices			
1-st	67.62%	75.09%	66.48%	74.08%
2-nd	17.84%	14.03%	17.77%	15.34%
3-rd	8.63%	6.79%	9.20%	6.70%
4-th	5.91%	4.08%	6.55%	5.18%

Table 13: Proportions of the best and preferred candidate labels selected by the reranker.

Furthermore, we rank the four candidate labels obtained via beam search according to their confidence and then analyze the distribution of the best labels and those preferred by our scorer. As illustrated in Table 13, in fewer than 70% of cases, the candidate label with the highest confidence is considered the best. In comparison, our scorer tends to favor candidates with higher confidence, highlighting areas for further improvement in this reranking step.