Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.047076

ARTICLE

Improve Chinese Aspect Sentiment Quadruplet Prediction via Instruction Learning Based on Large Generate Models

Zhaoliang Wu[1], Yuewei Wu[1,2], Xiaoli Feng[1], Jiajun Zou[3] and Fulian Yin[1,2,*]

[1]College of Information and Communication Engineering, Communication University of China, Beijing, 100024, China

[2]The State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, 100024, China

[3]Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China

*Corresponding Author: Fulian Yin. Email: yinfulian@cuc.edu.cn

Received: 24 October 2023 Accepted: 10 January 2024 Published: 26 March 2024

ABSTRACT

Aspect-Based Sentiment Analysis (ABSA) is a fundamental area of research in Natural Language Processing (NLP). Within ABSA, Aspect Sentiment Quad Prediction (ASQP) aims to accurately identify sentiment quadruplets in target sentences, including aspect terms, aspect categories, corresponding opinion terms, and sentiment polarity. However, most existing research has focused on English datasets. Consequently, while ASQP has seen significant progress in English, the Chinese ASQP task has remained relatively stagnant. Drawing inspiration from methods applied to English ASQP, we propose Chinese generation templates and employ prompt-based instruction learning to enhance the model's understanding of the task, ultimately improving ASQP performance in the Chinese context. Ultimately, under the same pre-training model configuration, our approach achieved a 5.79% improvement in the F1 score compared to the previously leading method. Furthermore, when utilizing a larger model with reduced training parameters, the F1 score demonstrated an 8.14% enhancement. Additionally, we suggest a novel evaluation metric based on the characteristics of generative models, better-reflecting model generalization. Experimental results validate the effectiveness of our approach.

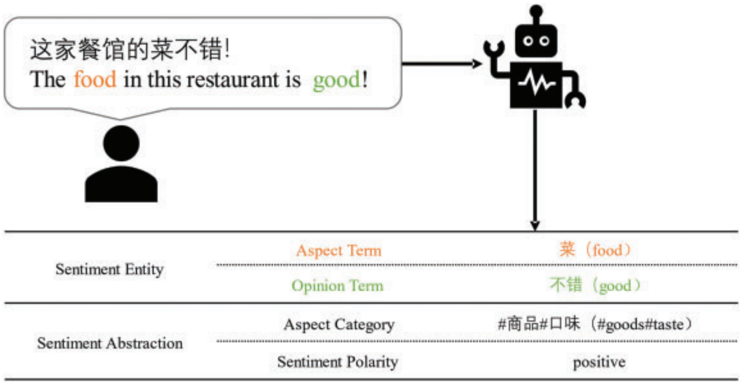
KEYWORDS

ABSA; ASQP; LLMs; sentiment analysis; Chinese comments

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) represents a specialized domain nested within sentiment analysis, a pivotal realm situated within the broader context of Natural Language Processing (NLP) and Machine Learning. Traditional sentiment analysis aims to determine the overall sentiment of a piece of text (e.g., positive, negative, or neutral), ABSA tasks employ a more fine-grained approach, focusing on specific aspects or attributes mentioned in the text and extracting one or multiple sentiment elements from the given text. Generally, these sentiment elements include the following four components: Aspect Term, Opinion Term, Aspect Category, and Sentiment Polarity [1]. Within this quartet of components, they can be divided into two major categories: 1) Sentiment Entities

(SE), encompassing opinion term and aspect term; 2) Sentiment Abstractions (SA), encompassing aspect category and sentiment polarity. Sentiment entities constitute spans existing within sentences, whereas sentiment abstractions are antecedent classifications inferred from the presence of sentiment entities. The sentiment elements are illustrated as shown in Fig. 1.



Sentiment Entity	Aspect Term	菜（food）
	Opinion Term	不错（good）
Sentiment Abstraction	Aspect Category	#商品#口味（#goods#taste）
	Sentiment Polarity	positive

Figure 1: An example of 4 sentiment elements of ABSA

In the preliminary phases of scholarly investigation, researchers directed their attention toward discrete sentiment elements including Aspect Term Extraction (ATE) [2–4], Aspect Category Detection (ACD) [5–7], and Aspect Sentiment Classification (ASC) [8–10]. Given the distinctive characteristics of opinion terms, their contextual relevance necessitates their alignment with specific aspects. Opinion terms devoid of corresponding facets render themselves devoid of substantive significance. Therefore, research on opinion terms is generally divided into two tasks: Aspect Opinion Co-Extraction (AOCE) [11,12] and Target-oriented Opinion Word Extraction (TOWE) [13,14]. Recent research endeavors have turned their attention toward the intricate realm of compound Aspect-Based Sentiment Analysis (ABSA), wherein the concurrent extraction of multiple interrelated affective constituents is pursued. This ambit encompasses binary sentiment elements extraction paradigms such as Aspect-Opinion Pair Extraction (AOPE) [15,16], End-to-End ABSA (E2E-ABSA) [17] and Aspect Category Sentiment Analysis (ACSA) [18] alongside more intricate ternary sentiment elements extraction frameworks like Aspect Sentiment Triplet Extraction (ASTE) [19,20] and Aspect Category Sentiment Detection (ACSD) [21,22]. The relationships between different ABSA tasks are depicted as shown in Fig. 2.

Despite the varying degrees of accomplishments observed across the aforementioned tasks within distinct domains, it is noteworthy that these endeavors could not provide a comprehensive aspect-level sentiment element representation. consequently, the task of extraction encompassing all four sentiment elements, known as Aspect Sentiment Quad Prediction (ASQP), remains a formidable challenge within this domain. The initial ASQP task utilized the TokenClass paradigm to classify each token within the text, aiming to ascertain spans (Sentiment Entities). Additionally, sentiment abstractions are affixed as complementary adjuncts to the labels [23]. In the wake of the evolution of large Language models (LLMs), researchers have forged innovative paradigms for ASQP tasks. They utilize these models in a Sequence-to-Sequence (seq2seq) manner to enhance the comprehension of textual content. This approach enables the models to generate sentiment quadruples that conform to natural language conventions [24]. In addition, a continuum of strategies for Parameter-Efficient Fine-Tuning (PEFT) [25–28] and low-precision parameter training [29] have emerged. These innovations have paved the way for the training of large language generative models, thereby accentuating the heightened importance of the seq2seq paradigm in contemporary research endeavors.

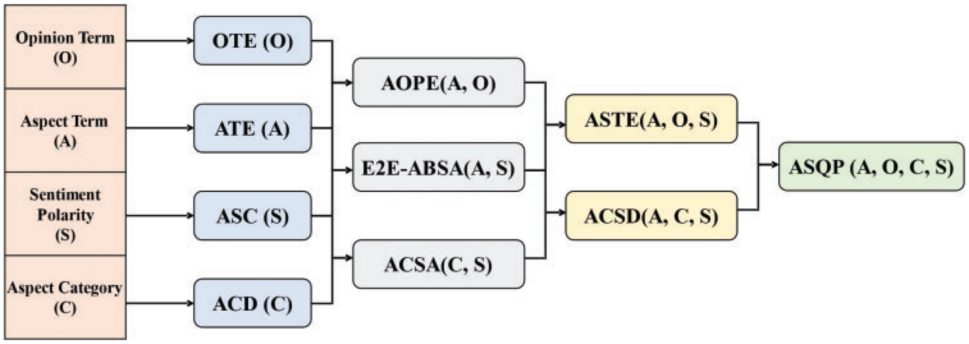


Figure 2: The relations between the four sentiment elements, single ABSA tasks, and compound ABSA tasks

Due to the exorbitant expenses associated with annotating an ABSA dataset, the majority of contemporary ABSA research is centered on linguistically affluent languages such as English. Nevertheless, a considerable cohort of researchers has effectively transposed their proficiency in ABSA tasks involving English texts to yield commendable results in the realm of Chinese scholarship. In recent times, the exploration of the Chinese ASQP task has been restrained owing to the paucity of Chinese quadruple datasets. However, remarkable strides have been made in the ASQP task grounded in English data. Researchers, who have delved into English data utilizing the seq2seq paradigm, have discerned that the sequencing of generated quadruples can exert a significant influence on the ultimate experimental outcomes. Consequently, they have augmented precision through the manipulation of the sequencing order [30] or by furnishing outcomes in multiple sequential orders [31]. Nonetheless, it is imperative to acknowledge that these methodologies have not undergone rigorous validation within the context of Chinese-language datasets.

Consequently, we are motivated by a keen academic interest to empirically assess the applicability and efficacy of these approaches within the domain of Chinese-language datasets. The results revealed that, when compared to English, Chinese displays a greater degree of variability, and the entropy associated with sentiment quadruple generation, based on pre-training models, exhibits a higher value. So, the change in entropy due to adjustments in the generation order of quadruples becomes less significant. Consequently, altering the sequence of sentiment quadruples can lead to certain improvements, but these effects are not particularly pronounced. In contrast, harnessing the linguistic knowledge embedded in pre-training models yields a more substantial enhancement in quadruple extraction, as these models have already assimilated extensive language-related knowledge during their pre-training. We assist the model in gaining a more comprehensive understanding of the meaning of sentiment elements by having it learn from various ABSA tasks, ultimately enhancing its performance on ASQP tasks.

In summary, the contributions of this work are three-fold:

- We conducted pilot experiments by experimenting with the transfer of generation templates that have been proven effective in English ASQP tasks to the Chinese language. This allowed us to select the most suitable ASQP generation templates for Chinese data.
- To fully exploit the implicit knowledge within large language models, we devised two methodologies: a data-reuse approach for multiple downstream tasks, and a training data augmentation

method utilizing additional ABSA datasets. Both strategies effectively enhance the training data without incurring additional annotation costs for the ASQP task, thereby elevating the model's performance.

- We have established a new evaluation metric that better reflects the model's generalization capabilities.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of relevant supplementary literature surrounding the Chinese ASQP task; Section 3 introduces the mainstream divisions of ASQP task paradigms; Section 4 outlines the pilot experiments we conducted; Section 5 presents our methodology; Section 6 discusses our dataset and experimental setup; Section 7 presents the final results and discussion; and Section 8 concludes the paper.

2 Related Work

Aspect-Based Sentiment Analysis (ABSA) is a critical subtask in Natural Language Processing (NLP), aimed at identifying the fundamental sentiment elements within a given sentence. Early researchers employed the TokenClass paradigm to annotate each token in the sentence, thereby transforming the initial task into a classification problem. Typically, researchers employ the BIOES tagging scheme [32] or its derivatives for intricate sentence annotation. This allows the model to categorize the tokens in the input sentence into a particular class, which is then compared and learned with the label. Usually, an encoder is used to map tokens into high-dimensional vectors, and subsequently, models like Conditional Random fields (CRF) [33] and Long short-term memory (LSTM) [34] serve as decoders to classify these vectors and obtain the final results. Liu et al. utilized word embeddings as the encoder component and employed a Recurrent neural network (RNN) as the decoder to extract opinion terms in sentences [35]. In parallel, Yin et al. [2] harnessed CRF, while Xu et al. [36] employed Convolutional Neural Networks (CNN) to address the same research objective. On the other hand, Wu et al. sought innovation in the realm of annotation. They propose an approach called Grid Tagging Scheme (GTS), a novel approach to annotating tokens that labels each token as either an opinion term, an aspect term, or neither. In addition, GTS also can identify opinion-aspect pairs, thereby facilitating the task of aspect-opinion pair extraction (AOPE) [19]. Additionally, some researchers have decomposed the primary task into two subtasks: entity extraction and relationship matching. They have introduced novel model architectures to optimize overall model performance. Zhao et al. introduced a structure known as Span-based multi-task learning (SpanMlt) [16]. Chen et al. proposed an innovative model containing dual-channel architecture [15]. However, when the ABSA task necessitates the extraction of an increased number of sentiment elements, the conventional annotation methods become inadequate. As a response, some researchers have adopted a pipeline approach, wherein two separate models are constructed to address distinct tasks. For instance, Peng et al. devised two Token Classification models to independently extract aspect-polarity pairs and opinion terms. Furthermore, they introduced a matching model to align the aforementioned two components and ultimately construct triplet predictions [37]. On the other hand, Xu et al. proposed the JET model, which makes the task of triplet extraction into a unified method by incorporating a position-aware tagging scheme [20]. Cai et al. took a step further by introducing two new datasets annotated with sentiment quadruples. They conducted benchmark tests on the task by constructing a series of pipeline baselines through the combination of existing models [23]. As research delves deeper, external knowledge has also been incorporated into models as an aid for sentiment element extraction. For instance, Wu et al. provided syntax knowledge as prior information to the model for sentiment analysis tasks [38]. While the aforementioned studies primarily focus on English data, they have served as significant inspiration for the development of sentiment analysis in the Chinese language. Peng et al. achieved promising results by leveraging the unique characteristics of the Chinese language to split sentences into radicals, characters, and words, conducting separate sentiment analyses on each level, and then integrating the results [39]. Yang et al. proposed a multi-task model that combined with domain-specific models, ultimately achieved state-of-the-art performance [40]. He et al. introduced the LGCF model, which is a multilingual learning model based on the interaction between local and global contextual focuses. It has demonstrated outstanding performance not only in Chinese datasets but also in English contexts [41]. Zhao et al. addressed the issue of aspect term omission in Chinese ABSA tasks by retraining word embeddings and validating the results on the dataset [42].

While the ABSA task was evolving, NLP was also rapidly advancing. Google pioneered the transformer architecture [43], and subsequently, they trained the BERT model using the encoder component of this architecture [44]. This marked the beginning of the era of pre-training NLP models that have had a profound impact on the NLP field. Researchers have observed that pre-training models, when fine-tuned with task-specific classifiers, exhibit remarkable adaptability and performance in the context of ABSA tasks. Sun et al. adeptly reformulated ABSA tasks as classification problems, harnessing the capabilities of the BERT model to enhance task performance effectively [45]. Other researchers have leveraged BERT's inherent capability—machine reading comprehension (MLC)—to address ABSA tasks [46–48]. In recent years, there has been a notable trend in the development of pre-training models, characterized by the continual expansion of model parameters and a distinct shift towards generative models. This trend is exemplified by several noteworthy contributions, including Google's T5 model [49], the llama model series proposed by Meta (formerly Facebook) [50,51], and the Generalized perturbation theory (GPT) series of models introduced by OpenAI [52,53]. Zhang et al. pioneered the transformation of the ABSA task into a seq2seq framework. They proposed the GAS (Generative Aspect-based Sentiment analysis) model, leveraging the fine-tuning of a generative language model to directly yield results in tuple format. Their methodology demonstrated notable efficacy across various subtasks within the ABSA domain [21]. In the same year, Zhang et al. further advanced their research by proposing the PARAPHRASE model, specifically designed to address the intricacies of the ASQP task. To better align the generative model with the task it encountered during pre-training, they implemented modifications to refine the GAS model's outputs, transforming them into structured natural language sentences. This strategic enhancement contributed to a notable performance boost in the context of the ASQP task [24]. Hu et al. took their research a step further by recognizing a crucial factor in the ASQP task – the sequential arrangement of the four sentiment elements from left to right significantly influenced the interdependencies among these elements within the sentiment tuple. This, in turn, had a substantial impact on the overall task performance. To address this issue, they employed pre-training models without fine-tuning to directly generate text. They systematically compared the generated results against different label orderings to identify the arrangement that minimized entropy. As a result, they introduced two novel models: the DLO (Dataset-Level Order) model, which accounted for order at the dataset level, and the ILO (Instance-Level Order) model, which considered order at the instance level. These models exhibited state of the art performance on different datasets, showcasing their effectiveness in the ASQP task [30]. In light of the research insights derived from previous studies, Guo et al. advanced their investigation by incorporating prompts into their model, facilitating the generation of results in diverse orders. They harnessed multiple generated outputs and employed methods such as voting or entropy analysis to ascertain the ultimate sentiment quadruple [31].

3 Foundational Concepts in Generative ASQP

For English data, two innovative yet distinct strategies are currently utilized within the framework of generative models for ASQP tasks: Paraphrase Generation [24] and Causal Deductive Generation [30,31].

3.1 Paraphrase Generation

Given a sentence x, the objective of the ASQP task is the comprehensive extraction of aspect-level quadruples $\{(a, o, c, s)\}$. In this paradigm, it is essential to transform the four elements of a quadruple into a sentence that conforms to common linguistic conventions. Regarding Sentiment Entities, since they originate from x, they inherently constitute words or phrases that adhere to common linguistic norms. However, Sentiment Abstractions, require certain transformations. 1) aspect category c is transformed into words, such as $x_c = {}'\text{service general}'$ for $c = {}'\text{service \# general}'$. 2) sentiment polarity $s \in \{positive, neutral, negative\}$ is mapped to words with sentiment semantics $x_s \in \{great, ok, bad\}$, respectively. It is noteworthy that certain datasets contain implicit aspect terms, necessitating the mapping of these implicit $a = null$ to $x_a = {}'it'$. Subsequently, these elements will be combined to form a natural language sentence, as illustrated in Eq. (1). If sentence x can extract multiple quadruples, they are separated by a special marker [SSEP].

$$x_c \; is \; x_s \; because \; x_a \; is \; x_o \tag{1}$$

Regarding pre-training generative language models, abstract summarization is an inherent aspect of their pre-training tasks. Consequently, the Paraphrase Generation ASQP emerges as a downstream task that exhibits enhanced learnability.

3.2 Causal Deductive Generation

The objective of the Causal Deductive Generation paradigm is in alignment with that of the Paraphrase Generation paradigm. However, it places greater emphasis on the fact that pre-training models are fundamentally generative models, whose goal is to rewrite the input sentence x into the target sentence y. The model operates sequentially over time steps, generating output based on the input x. Importantly, it leverages the previous output y_{t-1} as a contextual factor in the prediction process for the next time step y_t, persisting until the output reaches a designated termination marker, as shown in Eq. (2).

$$p\left(y_t\right) = f\left(x, y_0, \ldots, y_{t-1}\right) \tag{2}$$

where $f(\cdot)$ is the model, $p(\cdot)$ is the probability of each word in the vocabulary for its occurrence at time step t.

This paradigm accentuates the intrinsic causal relationships embedded within the generated sentences, thereby highlighting the pivotal role of the quadruple's arrangement order in shaping the model's performance. Leveraging the innate deductive capabilities inherent in pre-training models, organizing quadruples in congruence with the model's deductive reasoning aptitude can notably enhance results.

4 A Pilot Study

Under the elucidation provided in the preceding chapter, we conducted a pilot experiment to transfer the validated sentiment quadruple extraction technique, originally developed on English datasets, to a Chinese dataset. We replaced the original T5 model [49] with the mT0 model [54], which is adapted for Chinese, and simultaneously substituted the English words used in the PARAPHRASE Generation paradigm with their corresponding Chinese counterparts. Our results are presented in Table 1.

Table 1: Evaluation results, the best scores are marked in bold

Models	Precision (%)	Recall (%)	F1 (%)
GAS	46.64	48.44	47.51
PARAPHRASE	47.15	**51.17**	49.06
ILO	45.96	47.52	46.72
DLO(C-O-S-A) (1st)	47.33	49.60	48.41
DLO(C-S-O-A) (2nd)	46.99	49.75	48.33
DLO(A-O-S-C) (3rd)	47.73	51.04	49.30
DLO(S-A-C-O) (24th)	47.76	**51.17**	49.40
MvP	**51.09**	49.72	**50.39**

Here, the GAS model [21] is the initial attempt in the generation process; the PARAPHRASE model [24] renders the generated target into natural language form. DLO and ILO models [30] are Causal Deductive Generation models for generative models. They evaluate the entropy metrics by contrasting the pre-training model-generated outputs against various permutations of the label of the dataset. Notably, ILO prioritizes the selection of sentences based on their entropy scores, favoring those with the lowest entropy values as training sets. While DLO selects the permutation characterized by the minimal cumulative entropy score across all sentences for training purposes. In this pilot study, we chose to explore the lowest three permutations with the entropy values, as well as the permutation with the highest entropy value. The MvP model [31] is built upon the DLO model by incorporating additional prompt instructions. This allows the generative model to produce different permutations of sentiment quadrants for the same input x. Various techniques such as voting and entropy calculation are employed to select elements from different output variations, resulting in the final quadruples. Based on the pilot study results, we have the following observations.

The performance of these models on Chinese datasets exhibits patterns that differ from those on English datasets. For Chinese datasets, the PARAPHRASE model outperforms most other models. Its F1 score (49.06%) is higher compared to the subsequent SOTA models, ILO (46.72%) and DLO (48.41%). We have arrived at an initial inference that, concerning Chinese-language datasets, the generation of outputs adhering to the conventions of natural language is conducive to downstream tasks when utilizing generative language models.

The pre-training models without downstream fine-tuning do not yield effective results for Chinese-language datasets. However, the characteristic of altering the output sequence of sentiment elements to influence the model's performance is unchanged. We conducted downstream training on the pre-training model using templates based on sequences with the second and third lowest entropy values, as well as the highest entropy value. Comparing this approach to the DLO model (48.41%), it is noteworthy that the highest entropy sequence surprisingly exhibited the best performance (49.40%). We conducted more comprehensive supplementary experiments (see Appendix Table 1). The average entropy values for English datasets (Rest15 and Rest16 from 126 to 161) on the pre-training model

(T5 and mT0) were significantly lower compared to the Chinese dataset (405). Furthermore, the difference between the maximum and minimum entropy values in the English datasets (from 0.0295 to 0.0477) is an order of magnitude greater than that observed in the Chinese dataset (0.0043). We draw an empirical conclusion that in contrast to English, the Chinese language demonstrates a greater richness in morphological collocations, and its internal language sequencing displays a higher degree of variability.

Augmenting the prompts to guide the model in accomplishing analogous downstream tasks has shown a potential to enhance performance across those tasks. The MvP model currently represents the SOTA in performance on Chinese datasets. It achieves this by introducing prompts that guide the language model to generate templates with varying orders of sentiment elements. In our experiments, training was conducted using the top 5 sequences, and the final quad was determined through a voting mechanism.

5 Methodology

In light of the advancements in large language models, recent explorations in English ASQP tasks have pivoted towards the causal relations in generative models. This paradigm shift has culminated in a suite of methodologies centered on enhancing ASQP efficacy by refining the output sequencing of these models. However, as analyzed in the previous section, there exist significant disparities between ASQP in Chinese and English. Efficacious strategies in English contexts falter when transposed to Chinese datasets. In response to this divergence, our study probes into strategies tailored to the Chinese ASQP framework. Our initial step involved establishing a Chinese output paradigm by assessing loss values of prevalent templates in un-finetuned large language models. Subsequently, our research delved into leveraging identical datasets for a broad spectrum of downstream tasks, aiming to fully harness the latent knowledge embedded within the large Language models. This approach was further augmented by enriching the dataset with data from assorted ASBA sub-tasks. In our final analytical stride, acknowledging the multifaceted segment ability of Chinese entities, we introduced a quadripartite measurement standard grounded in Jessica's similarity metrics. The ensuing sections will offer an expansive discourse on these methodologies.

5.1 Output Template

The current trend in the development of generative models is moving towards multi-task dialogue models. These models generate responses y by understanding the input text x. So, we aim to design an output template that closely resembles natural language to minimize alterations to the internal knowledge of the original pre-training model. However, due to the necessity for batch processing of the model's outputs, a balance must be struck between conforming to natural language rules and adhering to predefined specifications. Inspired by Zhang et al.'s work, we have designed the output templates as follows:

$$y_{out} = (x_c \text{ 是 } x_s \text{ 因为 } x_a \text{ 是 } x_o) \tag{3}$$

where x_c and x_s are predefined categories, $x_a, x_o \in x$. "是" translates to "am," "is," or "are" in English, depending on the context while "因为" translates to "because". The utilization of this particular order of sentiment elements ($c \rightarrow s \rightarrow a \rightarrow o$) is rooted in its conformity to established natural language conventions. If a sentence contains multiple quads, they are separated by the "\t" symbol.

5.2 Multi-Task Enhanced Training

Differing from earlier models, large language pre-training models possess substantial parameter and training data volumes. Therefore, for them, multi-task training not only does not adversely affect performance but rather mutually enhances it. In parallel, this approach offers a potential solution to address the scarcity of Chinese ASQP datasets. Due to the comprehensive nature of the ASQP dataset in terms of sentiment quadrants, it becomes straightforward to replace it with other composite ABSA datasets. We repurpose the existing ASQP dataset by employing a prompt-based approach to train for six tasks: AOPE, E2E-ABSA, ACSA, ASTE, ACSD, and ASQP. Accordingly, based on each specific task, we prepend a task identifier at the beginning of the output to facilitate automatic recognition. We refer to this method as **Dataset Multi-Task Reuse (DMTR):**

$$x = x_{task} + x_{sentence} \tag{4}$$

$$y = y_{task} + y_{out} \tag{5}$$

$$y = LLMs(x) \tag{6}$$

where x_{task} and y_{task} respectively represent the task identifiers for input and output.

On the other hand, while the ASQP dataset may be limited, other Chinese composite ABSA task datasets can be employed in various tasks to achieve effective data augmentation. In our experiments, we chose to augment our data with those similar to the original dataset. Additionally, we selected comment data from different online platforms to assess the robustness of our experiments. It is worth noting that while both aspect category and sentiment polarity are considered sentiment abstractions and are selected from predefined sets, sentiment polarity tends to be more universal, with generally similar interpretations among different individuals. Aspect category, while, varies depending on the specific task and exhibits greater diversity. Therefore, for tasks involving the output of aspect categories, we refrained from training with other datasets. We refer to this method as **Dataset Multi-Source for Multi-Task (MSMT)**.

5.3 Tuning with LLMs

Fine-Tuning is one of the most common approaches to leveraging pre-training models for downstream tasks, driven by training data. In this experiment, we fine-tuned the mT0-base model. This model comprises 580 million parameters and was obtained through instruction tuning based on the t5 model. While it includes sentiment analysis in its instruction, it lacks the capability for ABSA tasks. Therefore, we fine-tuned it with minimizing the cross-entropy loss to meet our specific task requirements:

$$\mathcal{L}(x, y) = -log \sum_{t=1}^{n} (y_t | x, y_0, \ldots, y_{t-1}) \tag{7}$$

where n is the length of the output sequence y, as shown in Fig. 3.

Parameter-Efficient Fine-Tuning (PEFT), a method recently employed by researchers, enables the training of even larger-scale parameter LLMs. We employed one of these methods, known as Low-Rank Adaptation (LoRA) [25], to train the mT0-XL model. This model shares the same characteristics as the mT0-base but boasts an expansive parameter count of up to 3.7 billion. LoRA posits that the parameter changes required for adapting pre-training models to downstream tasks can be encapsulated within a low-rank matrix. Thus, it fixes the parameters of the original model and appends a low-dimensional full-rank matrix alongside it. Training this matrix facilitates driving the larger model effectively:

$$y = W_0 x + BAx \tag{8}$$

where W_0 represents the parameters of the original model, A represents the parameters that map inputs to a lower-dimensional space, and B represents the parameters that map back to the original dimension. B is initialized to zero, ensuring that the model is not affected by noise during the initial training phase, as shown in Fig. 4.

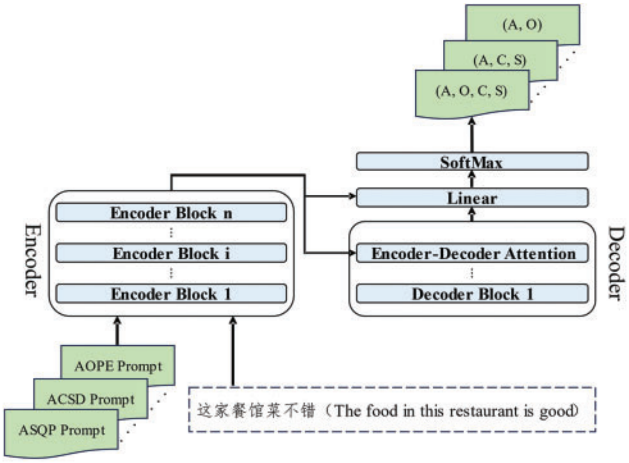


Figure 3: The training approach for LLMs based on fine-tuning

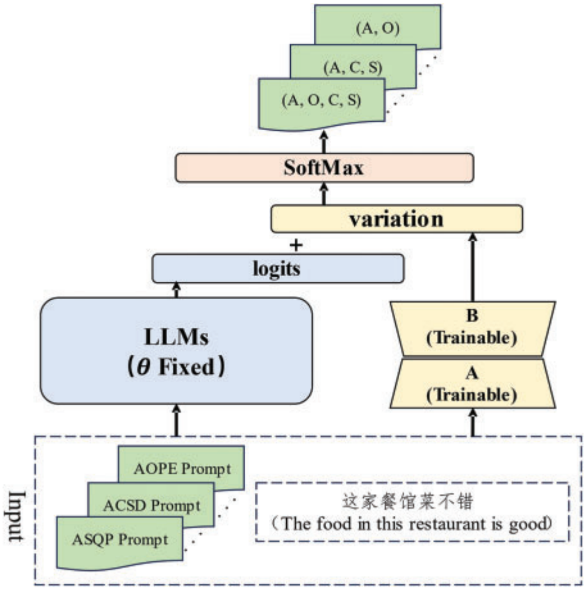


Figure 4: The training approach for LLMs based on LoRA

5.4 Evaluation Metrics

In previous studies, scholars typically regarded the F1 score as the primary evaluation metric for ASQP tasks. This metric is calculated by comparing the model's predicted sentiment quadruples with the labeled ones and deeming them correct if they match entirely:

$$precision = \frac{hit}{gold} \tag{9}$$

$$recall = \frac{hit}{predict} \tag{10}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{11}$$

where *hit* represents the number of correctly predicted quadruples, *gold* stands for the total number of annotated quadruples, and *predict* denotes the count of predicted quadruples. We also consider this metric as one of the references for our experimental results.

Although the F1 score serves as a valuable metric for classification tasks, it falls short in capturing the complexity of the ASQP task, which can be viewed as a compound task encompassing both recognition and classification components. In previous research, it has been customary to reformulate the recognition into a Token-Class paradigm, effectively rendering it a classification task. However, given the complexity of natural language, particularly when dealing with Chinese text, the addition or omission of a single character often has minimal impact on sentiment expression. This reality results in the absence of a standardized boundary delineation for aspect terms and opinion terms. Consequently, the extraction of these two sentiment elements entails a degree of ambiguity. For example, in the sentence "颗粒都很饱满 (The particles are all very plump)", the aspect term is "颗粒 (particles)". However, there can be different valid choices for the opinion term, such as "都很饱满 (all very plump)" or simply "很饱满 (very plump)". Inspiration from the prevailing use of generative models in solving ASQP has led us to novel metrics. We categorize quadruples into two groups: sentiment entities (aspect term and opinion term) treated as generative tasks using the Jaccard similarity as their text similarity measure; Sentiment abstractions (aspect category and sentiment polarity) are still treated as classification tasks, with successful matching indicating success and vice versa:

$$\text{TSCR} = \theta_1 \cdot jac\,(x_a, \widehat{x_a}) + \theta_2 \cdot jac\,(x_o, \widehat{x_o}) + \theta_3 \cdot Con(x_c, \widehat{x_c}) + \theta_4 \cdot Con(x_s, \widehat{x_s}) \tag{12}$$

$$jac\left(x, \hat{x}\right) = \frac{x \cap \hat{x}}{x \cup \hat{x}} \tag{13}$$

$$Con\left(x, \hat{x}\right) = \begin{cases} 1 & if\ x = \hat{x} \\ 0 & otherwise \end{cases} \tag{14}$$

where θ represents a tuning parameter, and by adjusting its value, we can enhance or diminish the importance of the corresponding sentiment element. In our experiments, this value was consistently set to $\frac{1}{4}$. TSCR stands for Text Similarity and Classification Results, a newly introduced metric in this experiment.

6 Datasets and Experiment Setups

6.1 Dataset

JD dataset: We collected reviews from the agricultural products category on the JD e-commerce platform and filtered them to meet our criteria. After subjecting each sentence to syntactic and semantic checks, we enlisted the assistance of three volunteers for annotation. In cases where all three annotators provided conflicting results, a fourth volunteer was consulted for the final decision. Ultimately, we annotated a dataset from JD that is comparable in scale to rest15, consisting of 780 samples for training and 260 samples for testing. The training set comprises 1,870 quadruples, while the testing set contains 652.

Meituan dataset: We collected restaurant review data from Meituan, a restaurant review platform in China. Similarly, after subjecting each sentence to syntactic and semantic checks, we enlisted the assistance of three volunteers for annotation. In cases where all three annotators provided conflicting results, a fourth volunteer was consulted for the final decision. We annotated a dataset from Meituan, comprising 3,000 samples. Unlike the JD dataset, this dataset is an ASTE dataset, containing only three sentiment elements: aspect term, opinion term, and sentiment polarity.

6.2 Implementation Details

In this experiment, we employed two models, mT0-base and mT0-XL, trained using fine-tuning and the LoRA method, respectively. Both models employed the AdamW optimizer, with initial learning rates of 1e-4 and 3e-4, respectively, and underwent decay over a specified number of epochs and steps. when the F1 score does not improve for ten consecutive epochs, it will trigger early-stop to stop training. We leverage five seed numbers for training including the main experiment, the comparison experiment. The obtained results are averaged as the final result.

7 Results and Discussion

7.1 Main Results

We adopt the models (GAS [21], PARAPHRASE [24], ILO [30], DLO [30], and MvP [31]) referenced in the pilot study as our baseline benchmarks, representing the SOTA across different periods in prior research. In addition, we conducted comparative experiments using ChatGPT[1]. We refined the prompt design[2] and employed both the gpt3.5-turbo and gpt4 engines, conducting tests in both 0-shot and 10-shot settings. In the DMTR experiment, we reused the JD dataset by adding prompts, enabling training for all text in AOPE, E2E-ABSA, ASTE, ACSA, ACSD, and ASQP tasks. Meanwhile, in the MSMT experiment, we randomly divided the Meituan dataset into three equal parts for training in AOPE, E2E-ABSA, and ASTE tasks, while the JD dataset was reused to participate in training for ACSA, ACSD, and ASQP tasks. The main results are reported in Table 2.

Our proposed approach demonstrates notable advancements when compared to previous state-of-the-art methodologies. With the utilization of the same pre-training model mT0-base, our DMTR method has achieved substantial improvements over the previously best-reported outcomes, demonstrating a 3.66% increase in Precision and a 5.92% increase in Recall. Moreover, there is a notable enhancement of 5.53% in the F1 score, along with a 2.00% rise in our self-defined metric, TSCR. Building upon the DMTR method, our MSMT approach has further elevated these metrics, yielding an additional increase of 0.11% in Precision, 0.46% in Recall, 0.26% in the F1 score, and a 0.86% improvement in TSCR. Upon employing a larger-scale Model, we observed a further amplification in our advancements. The DMTR and MSMT methods, in comparison to the previously optimal strategies, exhibited respective increments of 7.25% and 8.14% in the principal metric, the F1 score. However, it is pertinent to note that this enhancement predominantly stems from the augmented scale of the models. Consequently, our analysis primarily juxtaposes the improvements between the two methodologies (DMTR and MSMT) post-expansion. It is discernible that larger-scale models

[1] https://openai.com/chatgpt

[2] https://github.com/RidongHan/Evaluation-of-ChatGPT-on-Information-Extraction

contribute more significantly to improvements driven by diverse data sets. Additionally, ChatGPT, functioning as a conversational model, demonstrates robust performance in our metric, even when it cannot undergo downstream training and only relies on few-shot learning.

Table 2: Main results of the ASQP task on the Chinese dataset

	Precision (%)	Recall (%)	F1 (%)	TSCR (%)
GPT3.5-turbo (0 shot)	22.99	24.54	23.74	64.04
GPT3.5-turbo (10 shot)	36.34	42.64	39.24	70.67
GPT4 (0 shot)	32.36	37.27	34.64	70.87
GPT4 (10 shot)	41.49	51.23	45.85	74.65
GAS [21]	46.64	48.44	47.51	75.59
PARAPHRASE [24]	47.15	51.17	49.06	76.54
ILO [30]	45.96	47.52	46.72	68.66
DLO [30]	47.33	49.60	48.41	69.65
MvP [31]	51.09	49.72	50.39	71.26
DMTR (base)	54.75	57.15	55.92	78.54
MSMT (base)	54.86	57.61	56.18	79.40
DMTR (XL)	56.84	58.47	57.64	80.31
MSMT (XL)	**58.11**	**59.51**	**58.79**	**81.03**

7.2 Ablation Study

To further investigate the enhancement of the strategies we proposed for the Chinese ASQP task, we conducted a series of ablation experiments. The results are shown in Table 3. As previously described, DMTR utilized a single dataset but incorporated additional prompts to enable the model's multitask training on that dataset. Therefore, we conducted an ablation experiment by employing single-task training. In another ablation experiment, we employed the same input as MSMT, which is a multi-source dataset and multitask learning, but we modified the output format to a list structure. In the final results, it is evident that our proposed MSMT method achieved the best performance. However, different pre-training models exhibit varying paradigms across different metrics. When the model has fewer parameters, the output format has a significant impact on performance. However, as the model's parameter count increases, incorporating more data and engaging in multitasking learning can activate a broader range of knowledge within the pre-training model. It is worth noting that when the output format is in the list structure, both the base model and the XL model exhibit outstanding performance on the TSCR metric. The reason behind this outcome could be attributed to the larger volume of training data involved, which may have a more significant influence on the model. However, smaller-parameter versions of the model might struggle to comprehend this list-based output format, whereas the larger-parameter models have already grasped this format.

Table 3: Main results for the ASQP task and ablations using the proposed method

	Precision (%)	Recall (%)	F1 (%)	TSCR (%)
Single-task & sentence output (base)	53.47	57.30	55.32	78.25
Muti-task & list output (base)	53.19	54.94	54.05	78.94
DMTR (base)	54.75	57.15	55.92	78.54
MSMT (base)	**54.86**	**57.61**	**56.18**	**79.40**
Single-task & sentence output (XL)	56.16	59.11	57.59	79.38
Muti-task & list output (XL)	57.40	59.14	58.25	80.82
DMTR (XL)	56.84	58.47	57.64	80.31
MSMT (XL)	**58.11**	**59.51**	**58.79**	**81.03**

7.3 Discussions

Based on the results, it is evident that our proposed method is better suited for the ASQP task in Chinese datasets compared to the previous best-performing methods. With the advancement of language models, both the scale of models and training data continue to expand. Additionally, the use of instruction tuning has empowered large language models with enhanced generalization capabilities. GPT-3 and GPT-4 possess models with hundreds of billions and trillions of parameters, respectively. They possess the capability to perform ASQP tasks without undergoing traditional learning (parameter changes). However, when evaluated based on traditional metrics like F1 score, their performance on downstream tasks falls short compared to fine-tuning models in the hundred-million parameter. If we assess the performance based on our custom metric, which emphasizes the performance of generative models, the performance of GPT even surpasses some fine-tuning methods. Pre-training models, after undergoing instruction tuning, acquire enhanced reasoning abilities, thus demonstrating outstanding generalization capabilities even on unseen tasks. However, training them on downstream tasks inevitably leads to model changes, disrupting their generalization and resulting in catastrophic forgetting. This phenomenon becomes even more pronounced when dealing with complex languages like Chinese. Therefore, making the generation more natural and applying minor instruction tuning through the addition of prompts allows the model to retain its original characteristics to the fullest extent, ultimately enhancing its performance on ASQP tasks.

Simultaneously, by comparing the results obtained using the XL model and the base model, it can be observed that despite the significantly lower number of parameters required for training the XL model using the LoRA approach compared to the base model (shown in Table 4), its performance on ASQP subtasks is superior from various dimensions. This indicates that pre-training models have already acquired a sufficient amount of natural language knowledge and require appropriate methods to activate this knowledge to assist in accomplishing our desired downstream tasks.

Furthermore, we extracted the results from both the main experiment and the ablation experiments to count the mistakes for the four sentiment elements. These results have been visualized in Fig. 5. It can be observed that, for all four sentiment elements, consistent patterns are maintained across different pre-training models and methods. The extraction of opinion terms remains the most challenging aspect for the models. Given that sentiment polarity judgment is already a task during the mt0 model's pre-training phase, the model exhibits relatively accurate judgment regarding sentiment polarity. However, in the case of aspect category, mistakes in aspect term extraction can lead to classification mistakes, as there is a certain cause-and-effect relationship between them.

Table 4: The training approach for pre-training models and the trained parameters

	Training approach	Trainable params	All params	Trainable (%)
mT0-XL	LoRA	4.7 million	3.7 billion	0.126
mT0-base	Fine-tuning	582 million	582 million	100

Considering that a mistake in any element of a quad can result in a prediction mistake for the entire quad, we conducted further research based on the number of mistaken elements in a quad, and the results are presented in Fig. 6. In the case where all four elements in a quad are predicted incorrectly, it can be considered an invalid prediction. When three out of four elements in a quad are predicted incorrectly, it is primarily caused by errors in elements other than sentiment polarity. When three elements in a quad are predicted incorrectly, it is predominantly attributable to mistakes in elements other than sentiment polarity. In other cases, the patterns are similar to those depicted in Fig. 5, with opinion terms and aspect categories being responsible for the majority of mistakes. It is worth noting that when the pre-training model is set to mT0-base and the output format is in a list structure, there are instances where the model only outputs three elements. This indicates that language models with fewer parameters may struggle to fully comprehend the output format when the training data utilizes a list-based paradigm.

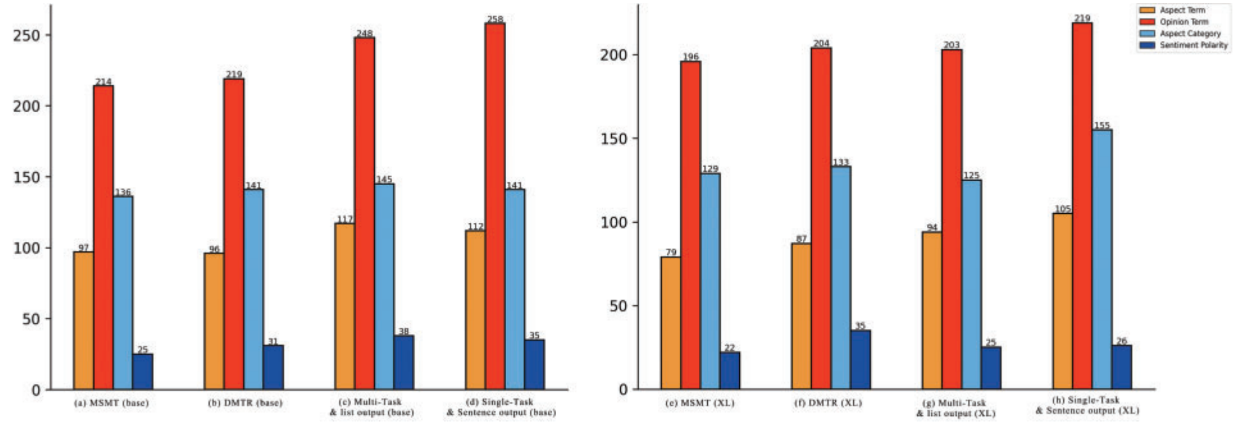
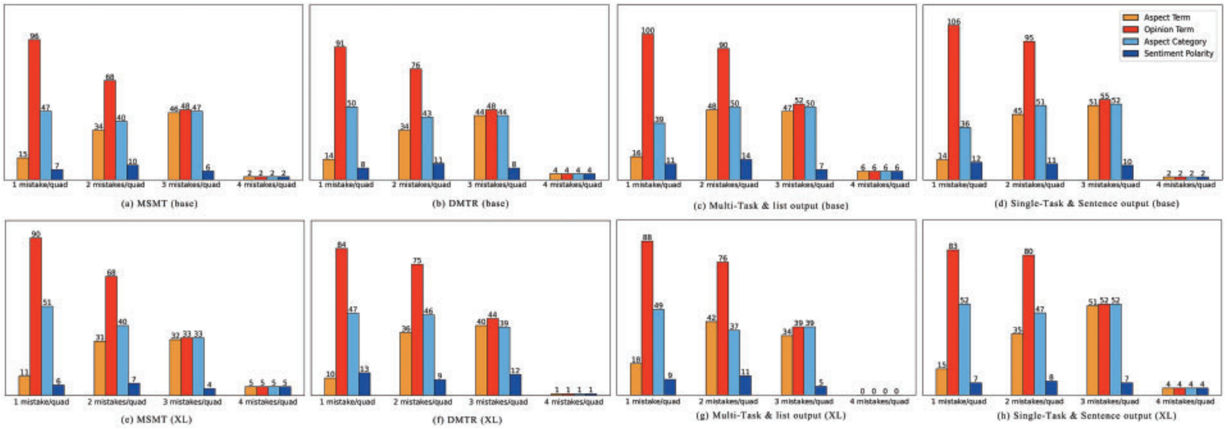


Figure 5: The number of mistakes for different sentiment elements in the results

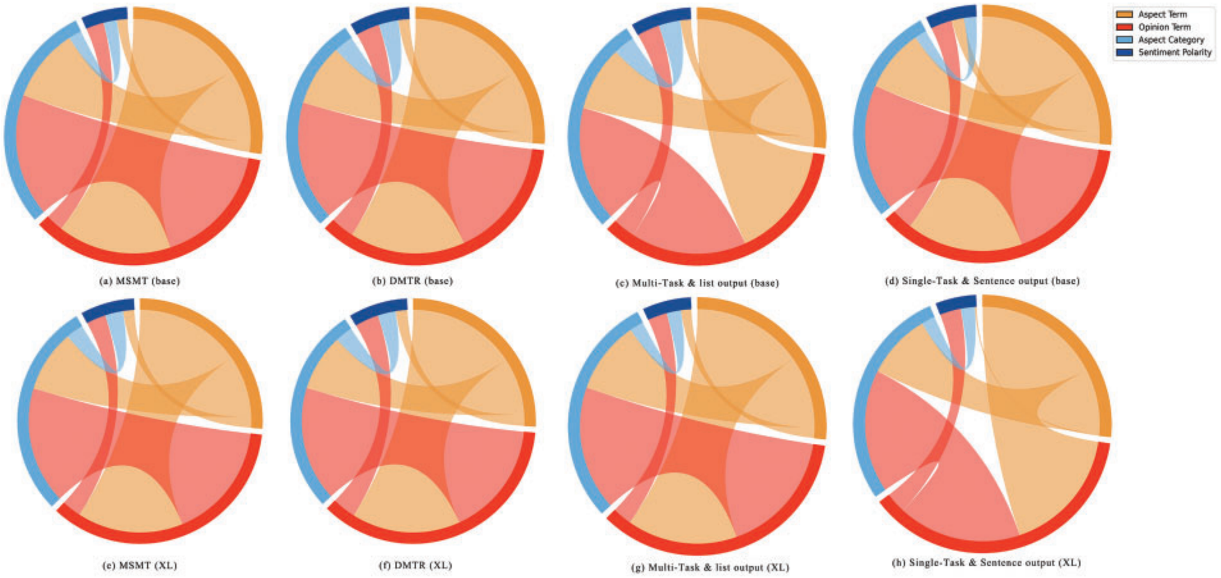
To further investigate the interplay between the four elements, we have plotted the co-occurrence of mistakes among them. In other words, we analyzed the frequency with which one element is incorrect when another element is also incorrect, as illustrated in Fig. 7. For the main results, they exhibit very similar patterns. The co-occurrence of mistakes in opinion terms with aspect terms and aspect category suggests that a mistake in extracting opinion terms can lead to associated mistakes in aspect term extraction. Moreover, a mistake in aspect term extraction can further impact aspect category classification, resulting in a cascade of mistakes. In the ablation experiments, the experiments using the mT0-base model with list output format and those using the mT0-XL model trained for a single task exhibited a similar paradigm. The reason for this phenomenon could be that both of them fail to facilitate the pre-training model's complete understanding of the downstream task, resulting in errors appearing in some invalid outputs.



(a) MSMT (base) (b) DMTR (base) (c) Multi-Task & list output (base) (d) Single-Task & Sentence output (base)

(e) MSMT (XL) (f) DMTR (XL) (g) Multi-Task & list output (XL) (h) Single-Task & Sentence output (XL)

Figure 6: The mistake for sentiment elements corresponding to the number of mistakes in each quad



(a) MSMT (base) (b) DMTR (base) (c) Multi-Task & list output (base) (d) Single-Task & Sentence output (base)

(e) MSMT (XL) (f) DMTR (XL) (g) Multi-Task & list output (XL) (h) Single-Task & Sentence output (XL)

Figure 7: Co-occurrence of mistake elements

7.4 Case Study

Table 5 shows two examples of our experiment. It can be observed from Example 1 that the phenomenon of polysemy in Chinese, can introduce linguistic ambiguity, thus leading to interpretational biases. In this example, the Chinese character "香" originally referred to any pleasant odor (fragrant), but as the language evolved, it also came to signify the taste of food being delicious (flavorful). In this sentence, the multiple meanings associated with this character introduce ambiguity. When labeling it, we opted for its original meaning, fragrant, but the model interpreted it as having a positive connotation related to "flavorful". In our JD dataset, the character "香" is quite common and predominantly used to describe the flavor of food. Unless there is explicit context indicating its meaning as “flavorful”, we interpret it as “fragrant”. However, whether through fine-tuning or LoRA, it remains challenging to completely overhaul the existing knowledge of large models in situations with such a limited amount of data. This also implies that the model possesses its own set of natural language understanding, and we need to guide it effectively to enhance its performance.

Table 5: In two cases, sentiment elements in underscores are incorrectly predicted

Example 1	
Sentence:	口感非常的好,特别好吃,特别香 The taste is very good, especially delicious and incredibly fragrant (flavorful)
Gold:	(口感, 非常的好, #商品#口感, 好), (null, 特别好吃, #商品#口味, 好), (null, 特别香, #商品#气味, 好) (Taste, very good, #goods#taste, good), (Null, especially delicious, #goods#flavor, good), (Null, incredibly fragrant, #goods#smell, good)
Predict:	(口感, 非常的好, #商品#口感, 好), (null, 特别好吃,特别香, #商品#口味, 好) (Taste, very good, #goods#taste, good), (Null, especially delicious and incredibly flavorful, #goods#flavor, good)
Example 2	
Sentence:	本来是准备送礼的,结果包装都破了,让我怎么送? It was originally intended as a gift but the packaging is all torn. How can I still give it as a gift?
Gold:	(包装, 都破了, #商品#包装, 差) (Packaging, all torn, #goods#packaging, bad)
Predict:	(包装, 破了,让我怎么送, #商品#包装, 差) (Packaging, torn. How can I still give it as a gift, #goods#packaging, bad)

In Example 2, the extraction of opinion terms has encountered the aforementioned issue of contentious entity boundary delineation. During annotation, we adhered to the principle of simplicity, selecting the simplest entity boundaries that could convey the author’s intent. However, the model’s selection of boundaries encompasses the reasons the author provides for generating sentiment. These two approaches do not have inherent superiority or inferiority; their effectiveness depends on the ultimate goal of the task. In our case, we aimed to leverage data-driven pre-training models to output the simplest entity boundaries, but this objective was not entirely achieved. In some instances, the model still produced outputs based on its pre-existing knowledge. However, this also underscores the importance of introducing new metrics because, in the context of natural language output, in Example 2, even an incorrect prediction can hold a significant reference value. While traditional metrics may classify this answer as outright incorrect, our proposed metric can assign some degree of score to it.

8 Conclusion

With the increasing emergence of large language models, harnessing their capabilities to better support our downstream tasks has become a focal point in current natural language processing research. As the parameters of large pre-training models continue to grow, cross-lingual models are poised to become mainstream. However, before reaching that point, understanding the characteristics of different languages and tailoring our approaches accordingly remains a valuable area of exploration. Our work is an attempt to enhance the Chinese ASQP task through the application of small-scale instruction learning. First, we select output patterns by comparing the performance across different output paradigms on the Chinese dataset. Second, we enhance the performance of the target task by introducing various types of auxiliary tasks. Finally, we establish a set of evaluation metrics that better align with the generative paradigm of ASQP. We achieved state-of-the-art results and also validated the applicability of the metrics we proposed.

As with most studies, there are still limitations that may guide future research directions. Firstly, both fine-tuning and LoRA pre-training models have encountered the issue of catastrophic forgetting, which could lead to the loss of internal knowledge within pre-training models, resulting in performance degradation. Secondly, while we have enhanced the model's understanding of sentiment analysis by introducing additional tasks, all tasks remain confined to ABSA. More diverse tasks could potentially stimulate the model's generalization capabilities, consequently improving the performance of ASQP tasks. Thirdly, our approach is entirely data-driven, however, due to limitations in the dataset size, it is challenging to enable the model to fully comprehend the partitioning rules. Therefore, incorporating some external rules might lead to improved results. It is worth noting that ASQP remains a challenging problem, particularly in languages with limited resources, warranting further exploration. In our forthcoming research endeavors, we aspire to leverage multi-turn dialogue models to extract diverse emotional elements through prolonged conversational interactions.

Acknowledgement: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team. Additionally, we extend our gratitude to ChatGPT for serving as a model in our comparative experiment.

Funding Statement: The work was supported by the National Key Research and Development Program (Nos. 2021YFF0901705, 2021YFF0901700); the State Key Laboratory of Media Convergence and Communication, Communication University of China; the Fundamental Research Funds for the Central Universities; the High-Quality and Cutting-Edge Disciplines Construction Project for Universities in Beijing (Internet Information, Communication University of China).

Author Contributions: Conceptualization, Zhaoliang Wu; Methodology, Zhaoliang Wu, Yuewei Wu; Data curation, Xiaoli Feng, Zoujia Jun; Formal analysis, Zhaoliang Wu; Investigation, Zhaoliang Wu, Fuliang Yin; Writing-original draft, Zhaoliang Wu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset can be provided on request. Experimental codes are available at: https://github.com/Wu00zl/MSMT.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," *IEEE Trans. Knowl. Data Eng*, vol. 35, no. 11, pp. 11019–11038, 2022. doi: 10.1109/TKDE.2022.3230975.

[2] Y. Yin, F. Wei, L. Dong, K. Xu, M. Zhang and M. Zhou, "Unsupervised word and dependency path embeddings for aspect term extraction," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, New York, NY, USA, 2016, pp. 2979–2985.

[3] X. Li and W. Lam, "Deep multi-task learning for aspect term extraction with memory interaction," in *Proc. 2017 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Copenhagen, Danmark, 2017, pp. 2886–2892.

[4] X. Li, L. Bing, P. Li, W. Lam, and Z. Yang, "Aspect term extraction with history attention and selective transformation," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Stockholm, SEW, 2018, pp. 4194–4200.

[5] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An unsupervised neural attention model for aspect extraction," in *Proc. 55th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Vancouver, Canada, 2017, pp. 388–397.

[6] L. Luo *et al.*, "Unsupervised neural aspect extraction with sememes," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, Macao, China, 2019, pp. 5123–5129.

[7] E. Ghadery, S. Movahedi, M. J. Sabet, H. Faili, and A. Shakery, "LICD: A language-independent approach for aspect category detection," in *Proc. 41st Eur. Conf. Inf. Retrieval (ECIR)*, Cologne, Germany, 2019, pp. 575–589.

[8] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention based LSTM for aspect-level sentiment classification," in *Proc. 2016 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Austin, TX, USA, 2016, pp. 606–615.

[9] W. Xue and T. Li, "Aspect based sentiment analysis with gated convolutional networks," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Melbourne, Australia, 2018, pp. 2514–2523.

[10] Y. Tay, L. A. Tuan, and S. C. Hui, "Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis," in *Proc. 32 AAAI Conf. Artif. Intell. (AAAI)*, New Orleans, LA, USA, 2018, pp. 5956–5963.

[11] J. Yu, J. Jiang, and R. Xia, "Global inference for aspect and opinion terms co-extraction based on multi-task neural networks," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 1, pp. 168–177, 2019. doi: 10.1109/TASLP.2018.2875170.

[12] M. Wu, W. Wang, and S. J. Pan, "Deep weighted maxsat for aspect-based opinion extraction," in *Proc. 2020 Conf. Empir Methods Nat. Lang. Process. (EMNLP)*, Punta Cana, Dominican Republic, 2020, pp. 5618–5628.

[13] A. P. B. Veyseh, N. Nouri, F. Dernoncourt, D. Dou, and T. H. Nguyen, "Introducing syntactic structures into target opinion word extraction with deep learning," in *Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Punta Cana, Dominican Republic, 2020, pp. 8947–8956.

[14] Z. Fan, Z. Wu, X. Dai, S. Huang, and J. Chen, "Target oriented opinion words extraction with target-fused neural sequence labeling," in *Proc. 17th Annual Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 2509–2518.

[15] S. Chen, J. Liu, Y. Wang, W. Zhang, and Z. Chi, "Synchronous double-channel recurrent network for aspect-opinion pair extraction," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Seattle, WA, USA, 2020, pp. 6515–6524.

[16] H. Zhao, L. Huang, R. Zhang, Q. Lu, and H. Xue, "SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Seattle, WA, USA, 2020, pp. 3239–3248.

[17] Z. Chen and T. Qian, "Relation-aware collaborative learning for unified aspect-based sentiment analysis," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Seattle, WA, USA, 2020, pp. 3685–3694.

[18] Z. Dai, C. Peng, H. Chen, and Y. Ding, “A multi-task incremental learning framework with category name embedding for aspect-category sentiment analysis,” in *Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Punta Cana, Dominican Republic, 2020, pp. 6955–6965.

[19] Z. Wu, C. Ying, F. Zhao, Z. Fan, X. Da and X. Rui, “Grid tagging scheme for aspect-oriented fine-grained opinion extraction,” in *Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Punta Cana, Dominican Republic, 2020, pp. 2576–2585.

[20] L. Xu, H. Li, W. Lu, and L. Bing, “Position-aware tagging for aspect sentiment triplet extraction,” in *Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Punta Cana, Dominican Republic, 2020, pp. 2339–2349.

[21] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, “Towards generative aspect-based sentiment analysis,” in *Proc. 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Joint Conf. Nat. Lang. Process. (ACL-IJCNLP)*, Bangkok, Thailand, 2021, pp. 504–510.

[22] C. Wu *et al.*, “Multiple-element joint detection for aspect based sentiment analysis,” *Knowl. Based Syst.*, vol. 223, no. 8, pp. 107073, 2021. doi: 10.1016/j.knosys.2021.107073.

[23] H. Cai, R. Xia, and J. Yu, “Aspect-category-opinion sentiment quadruple extraction with implicit aspects and opinions,” in *Proc. 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Joint Conf. Nat. Lang. Process. (ACL-IJCNLP)*, Bangkok, Thailand, 2021, pp. 340–350.

[24] W. Zhang, Y. Deng, X. Li, Y. Yuan, L. Bing and W. Lam, “Aspect sentiment quad prediction as paraphrase generation,” in *Proc. 2021 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Punta Cana, Dominican Republic, 2021, pp. 9209–9219.

[25] E. J. Hu *et al.*, “LoRA: Low-rank adaptation of large language models,” arXiv:2106.09685, 2021.

[26] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” in *Proc. 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Joint Conf. Nat. Lang. Process. (ACL-IJCNLP)*, Bangkok, Thailand, 2021, pp. 4582–4597.

[27] Q. Zhang *et al.*, “AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning,” arXiv:2303.10512, 2023.

[28] W. Zhen, R. Panda, L. Karlinsky, R. Feris, H. Sun and Y. Kim, “Multitask prompt tuning enables parameter-efficient transfer learning,” arXiv:2303.02861, 2023.

[29] Z. Feng *et al.*, “Towards unified INT8 training for convolutional neural network,” in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 1969–1979.

[30] M. Hu, Y. Wu, H. Gao, Y. Bai, and S. Zhao, “Improving aspect sentiment quad prediction via template-order data augmentation,” in *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Abu Dhabi, United Arab Emirates, 2022, pp. 7889–7900.

[31] Z. Gou, Q. Guo, and Y. Yang, “MVP: Multi-view prompting improves aspect sentiment tuple prediction,” in *Proc. 61th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Toronto, Canada, 2023, pp. 4380–4397.

[32] E. F. T. K. Sang and J. Veenstra, “Representing text chunks,” in *Proc.9th Conf. Eur. Chapter Assoc. Comput. Linguist. (EACL)*, Bergen, Norway, 1999, pp. 173–179.

[33] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proc. ICML*, Williamstown, MA, USA, 2001, pp. 282–289.

[34] G. Alex and A. Graves, “Long short-term memory,” *Supervised Sequence Labell. Recurr. Neural Netw.*, vol. 385, pp. 37–45, 2012.

[35] P. Liu, S. R. Joty, and H. M. Meng, “Fine-grained opinion mining with recurrent neural networks and word embeddings,” in *Proc. 2015 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Lisbon, Portugal, 2015, pp. 1433–1443.

[36] H. Xu, B. Liu, L. Shu, and P. S. Yu, “Double embeddings and cnn-based sequence labeling for aspect extraction,” in *Proc. 56th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Melbourne, Australia, 2018, pp. 592–598.

[37] H. Peng, L. Xu, L. Bing, F. Huang, W. Lu and L. Si, “Knowing what, how and why: A near complete solution for aspect-based sentiment analysis,” in *Proc. 34 AAAI Conf. Artif. Intell. (AAAI)*, New York, NY, USA, 2020, pp. 8600–8607.

[38] S. Wu, H. Fei, Y. Ren, D. Ji, and J. Li, "Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge," in *Proc. 30th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2021, pp. 3957–3963.

[39] H. Peng, Y. Ma, Y. Li, and E. Cambria, "Learning multi-grained aspect target sequence for Chinese sentiment analysis," *Knowl.-Based Syst.*, vol. 148, no. 15, pp. 167–176, 2018. doi: 10.1016/j.knosys.2018.02.034.

[40] Y. Heng, B. Zeng, J. Yang, Y. Song, and R. Xu, "A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction," *Neurocomput.*, vol. 419, no. 2, pp. 344–356, 2021. doi: 10.1016/j.neucom.2020.08.001.

[41] J. He, A. Wumaier, Z. Kadeer, W. Sun, X. Xin and L. N. Zheng, "A local and global context focus multilingual learning model for aspect-based sentiment analysis," *IEEE Access*, vol. 10, pp. 84135–84146, 2022. doi: 10.1109/ACCESS.2022.3197218.

[42] Q. Zhao, M. Zan, and M. Fan, "POS-ATAEPE-BiLSTM: An aspect-based sentiment analysis algorithm considering part-of-speech embedding," *Appl. Intell.*, vol. 53, pp. 27440–27458, 2023. doi: 10.1007/s10489-023-04952-3.

[43] V. Ashish *et al.*, "Attention is all you need," in *Proc. NeurIPS*, Long Beach, CA, USA, 2017, pp. 30.

[44] J. Devlin, M. W. Chang, L. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805, 2018.

[45] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proc. 17th Annu. Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 380–385.

[46] L. Gao, Y. Wang, T. Liu, J. Wang, L. Zhang and J. X. Liao, "Question-driven span labeling model for aspect-opinion pair extraction," in *Proc. 35 AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 12875–12883.

[47] S. Chen, Y. Wang, J. Liu, and Y. Wang, "Bidirectional machine reading comprehension for aspect sentiment triplet extraction," in *Proc. 35 AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 12666–12674.

[48] Y. Mao, Y. Shen, C. Yu, and L. Cai, "A joint training dual-MRC framework for aspect based sentiment analysis," in *Proc. 35 AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 13 543–13 551.

[49] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-totext transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.

[50] H. Touvron, T. Lavril, G. Izacard, X. Martinet, and M. Lachaux, "LLaMA: Open and efficient foundation language model," arXiv:2302.13971, 2023.

[51] H. Touvron, L. Martin, K. Stone, P. Albert, and A. Almahairi, "Llama 2: Open foundation and fine-tuned chat models," arXiv:2307.09288, 2023.

[52] A. Radford, J. Wu, R. Child, D. Luan, and D. Amodei, "Language models are unsupervised multitask learners," *OpenAI*, vol. 1, no. 8, pp. 9, 2019.

[53] T. Brown, B. Mann, N. Ryder, M. Subbiah, and J. Kaplan, "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[54] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, and S. Biderman, "Crosslingual generalization through multitask finetuning," in *Proc. 61th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Toronto, Canada, 2023, pp. 15991–16111.

Appendix

Table 1: The entropy value resulting from the outputs of the original pre-training model on the dataset. The maximum entropy values are indicated in bold, and the minimum entropy values are indicated with an underline

Order	JD_comments	Rest15 (T5)	Rest16 (T5)	Rest15 (mT0)	Rest16 (mT0)
A-O-C-S	405.56853	127.47517	126.05400	161.52647	159.15274
A-O-S-C	405.56728	127.47419	126.05293	161.52573	159.15204
A-C-O-S	405.56886	127.47741	126.05623	161.52782	159.15415
A-C-S-O	405.56887	127.51416	126.08039	161.57239	159.18354
A-S-O-C	405.56894	127.47451	126.05325	161.52643	159.15279
A-S-C-O	405.56962	127.51213	126.07822	161.57195	159.18309
O-A-C-S	405.56854	127.47519	126.05379	161.52574	159.15194
O-A-S-C	405.56852	127.47339	126.05189	161.52515	159.15132
O-C-A-S	405.56976	127.47582	126.05437	161.52609	159.15224
O-C-S-A	405.56922	127.47520	126.05375	161.52589	159.15202
O-S-A-C	405.56877	127.47532	126.05373	161.52573	159.15188
O-S-C-A	405.56926	127.47499	126.05341	161.52550	159.15164
C-A-O-S	405.56744	127.47789	126.05673	161.52804	159.15424
C-A-S-O	405.56842	**127.51523**	**126.08137**	**161.57286**	**159.18388**
C-O-A-S	405.56773	127.47692	126.05567	161.52750	159.15375
C-O-S-A	405.56701	127.47741	126.05617	161.52751	159.15375
C-S-A-O	405.56818	127.51412	126.08035	161.57243	159.18342
C-S-O-A	405.56712	127.47651	126.05542	161.52738	159.15365
S-A-O-C	405.57089	127.47388	126.05265	161.52615	159.15243
S-A-C-O	**405.57132**	127.51272	126.07879	161.57184	159.18296
S-O-A-C	405.57021	127.47399	126.05275	161.52599	159.15230
S-O-C-A	405.57112	127.47469	126.05340	161.52605	159.15231
S-C-A-O	405.57060	127.51174	126.07783	161.57134	159.18241
S-C-O-A	405.57033	127.47387	126.05255	161.52634	159.15266