BvSP: Broad-view Soft Prompting for Few-Shot Aspect Sentiment Quad Prediction

Yinhao Bai[1] **Yalan Xie**[2] **Xiaoyi Liu**[2] **Yuhua Zhao**[1] **Zhixin Han**[1]
Mengting Hu[1*] **Hang Gao**[3] **Renhong Cheng**[2]

[1] College of Software, Nankai University, [2] College of Computer Science, Nankai University,
[3] College of Artificial Intelligence, Tianjin University of Science and Technology
`{yinhao,xyl,xyiliu,zyh22,zhixinhan}@mail.nankai.edu.cn`,
`mthu@nankai.edu.cn`

Abstract

Aspect sentiment quad prediction (ASQP) aims to predict four aspect-based elements, including *aspect term*, *opinion term*, *aspect category*, and *sentiment polarity*. In practice, unseen aspects, due to distinct data distribution, impose many challenges for a trained neural model. Motivated by this, this work formulates ASQP into the few-shot scenario, which aims for fast adaptation in real applications. Therefore, we first construct a few-shot ASQP dataset (`FSQP`) that contains richer categories and is more balanced for the few-shot study. Moreover, recent methods extract quads through a generation paradigm, which involves converting the input sentence into a templated target sequence. However, they primarily focus on the utilization of a single template or the consideration of different template orders, thereby overlooking the correlations among various templates. To tackle this issue, we further propose a *Broad-view Soft Prompting* (BvSP) method that aggregates multiple templates with a broader view by taking into account the correlation between the different templates. Specifically, BvSP uses the pre-trained language model to select the most relevant k templates with Jensen–Shannon divergence. BvSP further introduces soft prompts to guide the pre-trained language model using the selected templates. Then, we aggregate the results of multi-templates by voting mechanism. Empirical results demonstrate that BvSP significantly outperforms the state-of-the-art methods under four few-shot settings and other public datasets. Our code and dataset are available at `https://github.com/byinhao/BvSP`.

1 Introduction

Analyzing user reviews, social media posts, product evaluations, and other content on the web to extract sentiment information related to specific aspects helps in understanding users' opinions and emotions regarding different aspects on the web. To monitor public opinion and support decision-making, the research field of sentiment analysis and opinion mining emerged (Vinodhini and Chandrasekaran, 2012; Shaik et al., 2022). The aspect sentiment quad prediction (ASQP) task aims to extract aspect quadruplets from a review sentence to comprehensively understand users' aspect-level opinions (Li et al., 2023; Hu et al., 2022). Recently, ASQP is gaining attention due to it involves predicting four fundamental aspect-level elements: 1) *aspect term* which is the concrete aspect description in the given text; 2) *opinion term* describing the exact opinion expression towards an aspect; 3) *aspect category* denoting the aspect type which refers to a pre-defined set, and 4) *sentiment polarity* indicating the sentiment class of the aspect. For example, given the sentence *"The room is clean."*, the aspect elements are "room", "clean", "room_overall", and "positive", respectively. Accordingly, the ASQP is described as a quad (*room*, *clean*, *room_overall*, *positive*).

However, in practical situations, aspect categories are not immutable and frozen (Zhou and Law, 2022). New aspects emerge as people discuss emerging phenomena, trends, products, and more through social media, news articles, and other means on the internet. As the restaurant domain illustrated in Figure 1, the initial aspect category set is pre-defined. Yet as the infrastructure upgrades, new aspects, such as *"WiFi"*, gradually appear. The sentence's category, i.e. *"internet"* does not exist in the pre-defined categories. This imposes challenges to the model's comprehensive and accurate understanding of the sentence. Moreover, the unseen aspect usually has a distribution shift, which is struggling for trained models to adapt accurately.

Therefore, researching the few-shot ASQP task, i.e. *fast adaptation to unseen aspects with only a few labeled samples*, becomes crucial, as it aligns more closely with real-world application scenar-

* Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8465–8482
August 11-16, 2024

Dataset	Sentence			Quad						Category	
	#S	#W	#W/S	EA & EO	IA & EO	EA & IO	IA & IO	#Q	#Q/S	#C	#M(C)
Rest15	1580	22886	14.48	1946	550	-	-	2496	1.57	13	192
Rest16	2124	30805	14.50	2566	729	-	-	3295	1.55	13	253
Restaurant	2284	34417	15.06	2431	530	350	350	3661	1.60	13	281
Laptop	4076	63879	15.67	3278	912	1241	342	5773	1.41	121	47
FSQP (Ours)	12551	149016	11.87	10749	151	5185	298	16383	1.31	80	205

Table 1: Data statistics and comparisons. #S, #W, #Q, and #C denote the number of sentences, words, quads, and categories, respectively. EA, EO, IA, and IO denote explicit aspect, explicit opinion, implicit aspect, and implicit opinion. #M(C) is the average number of instances in each category.

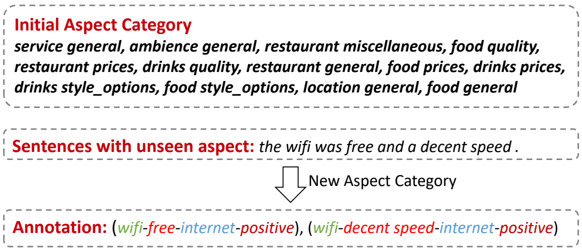


Figure 1: A unseen aspect case is shown. The newly emerged category *'internet"* is not mentioned in the pre-defined set of aspect categories.

ios. Yet, previous ASQP datasets either have a limited number of categories (Zhang et al., 2021a) or long-tailed distribution (Cai et al., 2021; Zhang et al., 2021a). This task lacks a proper benchmark dataset. Therefore, we annotate a few-shot ASQP dataset, named FSQP. This dataset aims to provide a more balanced representation and encompasses a wider range of categories, offering a comprehensive benchmark for evaluating few-shot ASQP.

Recent studies have employed generative methods to extract quads by converting input sentences into templated target sequences (Zhang et al., 2021b,a; Mao et al., 2022; Bao et al., 2022; Peper and Wang, 2022; Hu et al., 2022, 2023). Subsequently, by disentangling the formats of template, quads can be extracted. However, they have primarily concentrated on the utilization of a single template (Zhang et al., 2021b) or incorporate multiple templates by considering different quad orders (Hu et al., 2022), thereby ignore the correlation among these various templates. To overcome this limitation, we introduce an innovative method called Broad-view Soft Prompting (BvSP). BvSP leverages a pre-trained language model and utilizes Jensen-Shannon (JS) divergence to select several templates, enabling a more harmonious view of the available templates. We further introduce soft prompting to fine-tune the pre-trained language model with these selected templates. The final prediction is obtained from multiple templates by using a voting mechanism.

In summary, our major contributions of this paper are as follows:

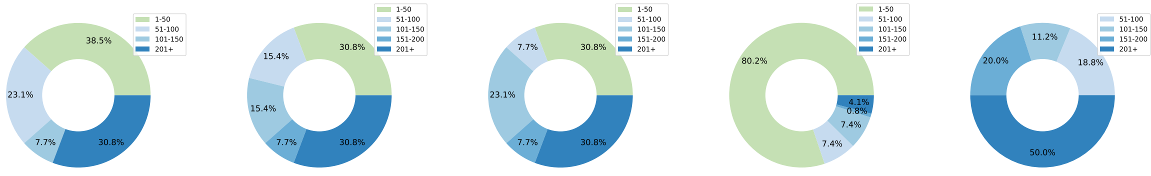
- We construct a new few-shot ASQP dataset FSQP which contains richer categories and is more balanced for the few-shot study. To the best of our knowledge, this is the first work to label the few-shot dataset in the ASQP task.
- We further propose BvSP, a various templates-based soft prompt learning method that improves quad prediction by taking into account the correlation between the different templates.
- Experimental results under four few-shot settings (i.e. one-shot, two-shot, five-shot, and ten-shot) demonstrate that BvSP outperforms strong baselines and has significant gains in other public datasets.

2 Datasets

To construct a dataset that is more representative of real-world scenarios, we manually annotate a new dataset, named **F**ew-**S**hot AS**QP** (FSQP). In this section, we will first describe the data collection and annotation process. Following that, we will establish the superiority of FSQP by comparing it with previous datasets in terms of key statistics and features.

2.1 Collection

Our data source for this study is a collection of Yelp reviews spanning six years and originating from diverse cities across the United States. These reviews encompass various establishments, including restaurants, hotels, and beauty spas. Initially, these reviews were labeled with aspect categories and



(a) Rest15 (b) Rest16 (c) Restaurant (d) Laptop (e) FSQP

Figure 2: The category distribution is presented according to the number of instances. For example, the green section indicates the proportion of categories with the number of instances between 1 and 50.

sentiment polarity by Bauman et al. (2017). The FSQP dataset is an extension of this Yelp review data, featuring additional annotations and refinements.

2.2 Annotation

2.2.1 Annotation Guidelines

We consolidate the assessments across the three domains by aligning them with the same categories as per the evaluations carried out in prior works (Bauman et al., 2017; Li et al., 2019). These categories encompass aspect categories and sentiment polarities, initially identified by the Opinion Parser (OP) system (Qiu et al., 2011a; Liu, 2020). The annotations generated by this system are manually double-checked for accuracy and completeness. Afterward, we perform a thorough examination of the taxonomy and merge aspect categories that share similarities. Building on the double-checked aspect categories and sentiment polarity, we proceed to enrich the annotations of the *aspect term* and the *opinion term*.

Additionally, in line with the approaches proposed by Cai et al. (2021); Poria et al. (2014), we also consider the annotation of implicit *aspect terms* and implicit *opinion terms*.

2.2.2 Annotation Process

For selecting the aspect categories, the decision is made by a team of professional annotators. They check each category and its similarity one by one. For the annotation of other elements, two master's students who are well-versed in aspect-based sentiment analysis are chosen as annotators to annotate independently. The strict quads matching F1 score between two annotators is 78.63%, indicating a substantial agreement between them (Kim and Klinger, 2018). If one of the annotators disagrees with any content of the quad, they discuss to reach a consensus. Meanwhile, the leader of the master students will help to make the final decision. Throughout the six-month annotation period, while annotators were allowed to communicate with each other, their annotation operations remained relatively independent. Following this, there was a two-month period specifically set aside for verifying the accuracy of the annotations, during which we rigorously enforced consistency. Based on the above measures, the annotation of the new dataset has undergone careful scrutiny and discussion, suggesting a high level of credibility.

2.3 Statistics and Analyses

The statistics of FSQP can be found in Appendix §D. FSQP comprises 12,551 sentences, yielding a total of 16,383 quads. It's important to note that FSQP also encompasses implicit information, specifically, *implicit aspect terms* and *implicit opinion terms*, which are not explicitly mentioned in the sentence. Upon comparing these two types of implicit information, it becomes evident that the implicit opinion terms are more numerous than the implicit aspect terms.

In Table 1, we proceed with a more detailed comparison between FSQP and existing ASQP datasets regarding their distribution. It's evident from the table that FSQP surpasses even the current largest benchmark dataset, Laptop, in terms of both scale and the number of quads. As a result, FSQP **stands out as a dataset with a more extensive collection of instances and quads.** Otherwise, based on the statistics of categories, it can be found that the existing benchmark dataset has insufficient categories and instances in each category. Our dataset effectively addresses both of these shortcomings. FSQP has a larger number of categories than the restaurant domain. Compared to Laptop, it has a higher average number of instances in the category.

The category distribution is also counted according to the number of instances in a certain interval. As shown in Figure 2, it can be seen that the categories with only 1-50 instances in previous bench-

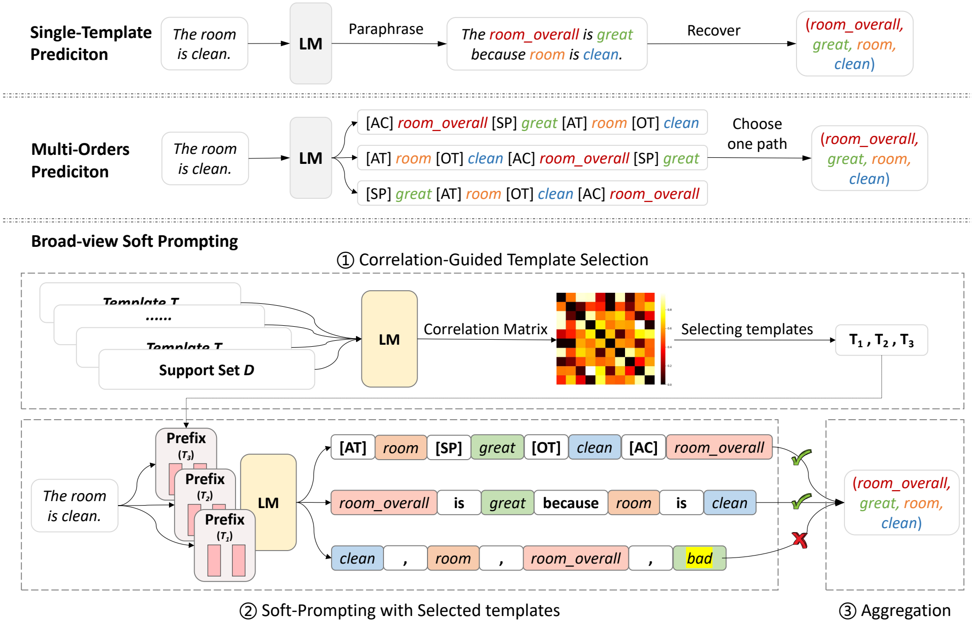


Figure 3: An overview of the proposed Broad-view Soft Prompting (BvSP). The single-template prediction is Paraphrase (Zhang et al., 2021b). The multi-order prediction approach is DLO (Hu et al., 2022). BvSP combines these templates as candidates and proposes a correlation-guided strategy for template selection.

mark datasets account for a majority percentage. Especially in the Laptop, such categories account for 80.2%. On the contrary, FSQP rarely has tailed categories. **Therefore, stable training and reliable testing can be achieved on FSQP.** In summary, FSQP provides a suitable testbed to identify the capabilities of ASQP methods in real-world scenarios. It is worth noting that the design of FSQP follows the FewRel dataset (Han et al., 2018), which is proposed for few-shot relation classification. FewRel has a wide range of relation classes and balanced distribution, with 100 relations and each relation contains 700 instances.

3 Methodology

3.1 Formulation and Overview

Given a sentence x and its aspect sentiment quads $\{(at, ot, ac, sp)\}$. Following the previous generation-based works (Zhang et al., 2021a; Hu et al., 2022), we define projection functions to map the quads (at, ot, ac, sp) into semantic values $(x_{at}, x_{ot}, x_{ac}, x_{sp})$. For example, we map the *"POS"*, *"NEU"*, and *"NEG"* labels of sentiment polarity to *"great"*, *"ok"*, and *"bad"*, and map the *"NULL"* label of aspect term to *"it"*. Based on the above rules, we use several templates, as shown in Figure 7, to convert the aspect sentiment quadruples into target sequences that the language model can understand. If a sentence contains multiple quads, the target sequences are concatenated with a special marker [SSEP] to obtain the final target sequence.

As shown in Figure 3, we first use the pre-trained language model with JS divergence to select appropriate subsets of these templates, considering the efficiency and effectiveness of optimization, for a more detailed explanation, please refer to §4.6. Conditioned on soft prompts for different templates, it can generate multiple quads from diverse views. During the inference phase, different templates are generating results from different views. Thus, we aggregate these templates by voting for accurate quads extraction. Next, we will describe each section in detail.

3.2 Correlation-Guided Template Selection

Though there are several available templates, jointly using them is inefficient. Thus, we choose templates by evaluating their correlations with a pre-trained language model (LM), where the detailed pre-training process is described in §4.1. In this way, chosen templates are more suitable for the nature of LM. Concretely, the correlation between the two templates is based on the whole support set $\mathcal{D}$. We compute the average score across all instances of these two templates. Specifically, given an input x, its quads, and several available templates $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_T\}$ where T is the number

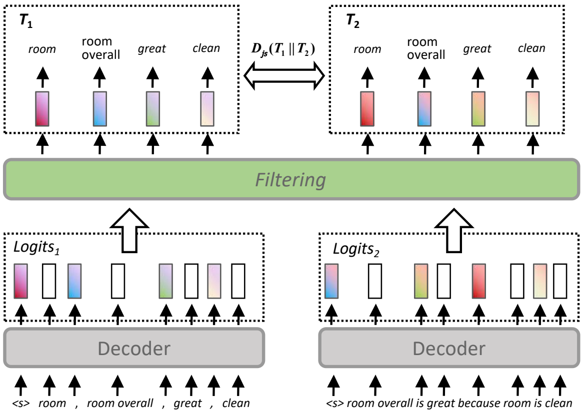


Figure 4: Quad elements (colored ones) of the output sequence from the decoder are filtered, which are leveraged for computing template correlations.

of templates, we use $\boldsymbol{y^i}$ and $\boldsymbol{y^j}$ to indicate the target sequences under any two templates $\mathcal{T}_i$ and $\mathcal{T}_j$. The target sequence is usually composed of quad elements and linking symbols. As shown in Figure 4, for both two templates, (*"room"*, *"room_overall"*, *"great"*, *"clean"*) denote the quad elements. The *","*, *"is"*, and *"because"* indicate symbols linking the quad.

Therefore, we can choose the quad part from each target sequence. We denote this process by the following equation.

$$\boldsymbol{h^i} = \text{Filtering}(\boldsymbol{x}, \boldsymbol{y^i}) \tag{1}$$

where $\boldsymbol{h^i}$ is set of selected representations from the target sequence and $\boldsymbol{y}^i$ is fed into the decoder as teacher forcing (Williams and Zipser, 1989). Then, we obtain the correlation of two target sequences by using JS divergence.

$$\begin{aligned}\mathcal{D}_{KL}(\boldsymbol{p}||\boldsymbol{q}) &= \sum_{|V|} \boldsymbol{p}\log(\frac{\boldsymbol{p}}{\boldsymbol{q}})\\ \mathcal{D}_{JS}(\boldsymbol{h^i}||\boldsymbol{h^j}) &= \frac{1}{|\boldsymbol{h^i}|}\sum_{|\boldsymbol{h^i}|}(\frac{1}{2}\mathcal{D}_{KL}(\boldsymbol{h^i}||\frac{\boldsymbol{h^i}+\boldsymbol{h^j}}{2})\\ &\quad +\frac{1}{2}\mathcal{D}_{KL}(\boldsymbol{h^j}||\frac{\boldsymbol{h^i}+\boldsymbol{h^j}}{2}))\end{aligned} \tag{2}$$

where $|V|$ is the size of the vocabulary set and $|\boldsymbol{h^i}|$ is the number of selected quad tokens. The $\mathcal{D}_{KL}$ represents the calculation of the KL divergence of two probability distributions.

For the whole support set $\mathcal{D}$, we have $(\boldsymbol{x}, \{\boldsymbol{y}^i\}_{i=1}^T)$ for each instance by constructing templates. Then the average score of templates is calculated over the support set:

$$\mathcal{S}_{\mathcal{T}_i,\mathcal{T}_j} = \frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}} \mathcal{D}_{JS}(\boldsymbol{h^i}||\boldsymbol{h^j}) \tag{3}$$

where $\mathcal{S}_{\mathcal{T}_i,\mathcal{T}_j}$ denotes the average correlation of all instances between templates $\mathcal{T}_i$ and $\mathcal{T}_j$.

Then with the correlation between every two templates, we can obtain a correlation matrix $\mathcal{S} \in \mathbb{R}^{|T|\times|T|}$. Then, we enumerate the entire matrix and find the k matrix points with the smallest values. These points are the most relevant k templates while are our final fine-tuned templates. We aggregate multiple templates by taking into account their shared quads. When the templates exhibit greater disparity, conflicts arise, leading to inconsistent quads and empty predictions. Conversely, when the correlation between templates is stronger, they are more harmonious, fostering consistent support for the quad.

3.3 Soft-Prompting with Selected Templates

Then we aim to incorporate various selected templates. Yet it is difficult to distinguish between them. Thus we design a specific prefix for each template that can act as an indicator for each of them. With the input $\boldsymbol{x}$, the target sequence $\boldsymbol{y}_t$, and the succession of prefix parameters $\boldsymbol{z^i}$ for the template $\mathcal{T}_i$ (Li and Liang, 2021), we fine-tune with minimizes cross-entropy loss defined as follows:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y^i}) = -\sum_{t=1}^{n} \log p_\theta(\boldsymbol{y}_t^i|\boldsymbol{x}, \boldsymbol{z^i}, \boldsymbol{y}_{<t}^i) \tag{4}$$

where n is the length of the target sequence $\boldsymbol{y^i}$.

3.4 Multi-Templates Aggregation

During the inference phase, following (Gou et al., 2023), BvSP aggregates the quad results generated by all selected templates. Subsequently, we employ a voting mechanism to determine the quad that garners the high frequency of predictions and designate it as the final prediction. By defining a threshold τ, the quad is extended to the final prediction when the number of votes for this quad is larger than this threshold.

$$\mathcal{P} = \{q|q \in \bigcup_{i=1}^{k} \mathcal{T}_i \text{ and } (\sum_{i=1}^{k} \mathbb{1}_{\mathcal{T}_i}(q) \geq \tau)\}$$

where q denotes the quad obtained in the template $\mathcal{T}_i$ and $\mathcal{P}$ is the final prediction of quads.

4 Experiments

4.1 Implementation Details

In the experiments, all the reported results are the average of 5 runs. We adopt T5-base (Raffel et al.,

Baseline	One-shot			Two-shot			Five-shot			Ten-shot			Avg		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
GAS	<u>26.39</u>	<u>26.70</u>	<u>26.54</u>	40.42	41.21	40.81	49.70	49.89	50.50	51.89	52.80	52.34	42.10	42.65	42.55
Paraphrase	23.53	23.25	23.38	39.04	38.92	38.98	49.04	49.54	49.29	52.02	52.52	52.26	40.91	41.06	40.97
DLO	26.37	26.14	26.23	36.70	38.80	37.72	49.05	**<u>51.80</u>**	50.38	50.48	52.83	51.62	40.65	42.39	41.48
ILO	25.47	24.35	24.98	37.71	38.54	38.12	<u>50.64</u>	51.53	<u>50.95</u>	50.52	53.04	51.94	41.08	41.87	41.50
MvP	25.82	25.86	25.84	<u>40.64</u>	<u>41.65</u>	<u>41.14</u>	49.05	50.43	49.73	<u>53.05</u>	**54.18**	<u>53.61</u>	<u>42.14</u>	<u>43.03</u>	<u>42.58</u>
ChatGPT	23.79	23.24	23.56	33.24	38.14	35.52	36.29	40.63	38.97	46.03	40.25	42.96	34.84	35.56	35.25
BvSP	**45.50**	**32.62**	**37.99**	**52.12**	**43.41**	**47.34**	**56.87**	50.55	**53.52**	**57.29**	52.04	**54.53**	**52.95**	**44.66**	**48.35**

Table 2: Evaluation results of few-shot ASQP task, compared with baseline methods in terms of precision (`Pre`, %), recall (`Rec`, %) and F1 score (`F1`, %). Under each column, the best results are marked in bold and the best baseline results are underlined.

2020) as the pre-trained generative model. For all baselines and our method, they are pre-trained first on the training set and then fine-tuned with the support set. Our work follows the common few-shot setting, where regardless of the number of shots, all samples from the training set are used to train the model. The *"k-shot"* concept refers to the number of labeled instances provided for each class during the few-shot task. For example, in a 1-shot setting, only one labeled instance is available per class in the support set, while in a 2-shot setting, two labeled instances are available per class. In our experiments, for k-shot, we randomly sample k instances in the test set as the support set and the remaining instances in the test set as the query set. Especially, BvSP uses all templates to pre-train the prefix parameters on the training set. It is worth noting that in BvSP we freeze the LLM parameters and only fine-tune the parameters of the prefix.

In addition, we also depict the template details of each baseline, more details about the templates are given in Appendix §B. During the pre-training phase, we set the epoch to 20, batch size to 16, and learning rate to 3e-4, except for DLO, ILO, and MvP, where the learning rate is set to 1e-4. For the support sets in the fine-tuning stage, we maintain the epoch at 20, choose a batch size of 8, and utilize a learning rate of 3e-4. During the inference stage, all methods employ a beam size of 1, except for DLO and ILO, which use a beam size of 5.

4.2 Datasets

The original dataset `FSQP` is first divided into a training set, a development set, and a testing set according to aspect category. More details about datasets are given in Appendix §A. Few-shot learning aims to replicate real-world scenarios where the model encounters novel classes not included in the training dataset. Consequently, these three datasets have varying numbers of aspect categories with no overlap. We further conduct the experiments on our dataset `FSQP` under the four few-shot settings.

4.3 Compared Methods

We choose the strong generative baseline methods. They include both the currently most popular LLMs, i.e. ChatGPT (Brown et al., 2020), as well as the state-of-the-art (SOTA) methods in sentiment analysis in recent years, namely GAS (Zhang et al., 2021b), Paraphrase (Zhang et al., 2021a), DLO, ILO (Hu et al., 2022), and MvP (Gou et al., 2023).

4.4 Experimental Results

Experimental results of the ASQP task under few-shot settings are reported in Table 2. For a fair comparison, we follow the settings of Hu et al. (2022) and set the default number of selected templates to top-3. Our method BvSP still has room for improvement in the performance (see §4.6).

It can be seen that BvSP achieves the best performance in the four few-shot settings. Especially, compared with baseline GAS, BvSP gains absolute F1 score improvement by +11.45% under the one-shot setting. Moreover, BvSP is able to outperform DLO and ILO with the same number of selected templates under the same few-shot settings. Under the same few-shot settings, BvSP demonstrates superior performance compared to MvP. These results validate the effectiveness of BvSP in providing a broader view of templates.

	Method	Ten-shot		
		Pre	Rec	F1
	BvSP (JS Min)	**57.29**	**52.04**	**54.53**
I	BvSP (JS Max)	56.29	50.09	53.01
	BvSP (Entropy Min)	62.90	44.30	51.98
	BvSP (Entropy Max)	61.42	45.67	52.19
	BvSP (random)	56.27	50.27	53.10
II	BvSP (rank)	56.29	50.09	53.01
	BvSP (rand)	54.27	51.38	52.78

Table 3: Evaluation results of ablation study.

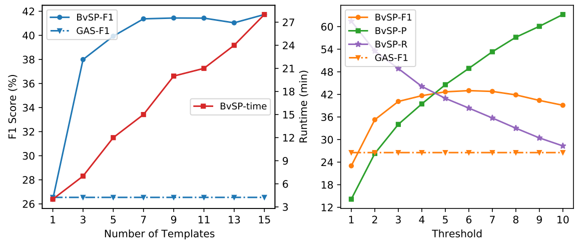


Figure 5: Effects of hyperparameters on FSQP under the one-shot settings.

4.5 Ablation Study

To demonstrate the effectiveness of BvSP in selecting templates and aggregation methods, ablation experiments are performed. The results are shown in Table 3. Following the default setting of BvSP, the model variants also select top-3 templates. The model variants first study from (**I**) template selection strategy, including maximum and minimum entropy, maximum JS divergence, i.e. BvSP (JS Max), random sampling, i.e. BvSP (random). During inference, we further study the (**II**) aggregation strategies, including BvSP (rank) and BvSP (rand). The former selects the top-ranked sequence by considering the perplexity of the generated sequences, while the latter selects one sequence randomly.

It is first observed that using the minimal JS divergence consistently outperforms the other strategies. Specifically, compared with BvSP (JS Max), BvSP (Entropy Min), BvSP (Entropy Max), and BvSP (random), BvSP makes absolute F1 score improvements by +1.52%, +2.55%, 2.34%, and +1.43%, respectively. These results underscore the effectiveness of our strategy. Selecting correlated templates can make them cooperate harmoniously.

Furthermore, BvSP outperforms BvSP (rank) and BvSP (rand) by +1.52% and +1.75% F1 score. Our aggregation method using voting demonstrates superiority over the remaining two strategies. Actually, both of these strategies produce outputs from a single template, which could be either random or ranking selection. In contrast to these methods, we select the final prediction by considering the commonalities between multiple templates.

4.6 Hyperparameter Study

The study includes an examination of the effects of two hyperparameters: k and τ. The k represents the number of selected templates, while τ refers to the threshold for gaining final quads from multiple templates. Figure 5 illustrates the resulting curves.

Number of selected templates k: This hyperparameter affects the breadth of templates view. Here we set various k values and also change the threshold $\tau = k/2$ accordingly. All other hyperparameters are kept the same. Analyzing the left plot of Figure 5, it is seen that BvSP outperforms the baseline GAS in most cases, indicating the robustness of this hyperparameter to some extent. While a large k increases time consumption, opting for a smaller value results in reduced performance. The figure clearly demonstrates that when considering both the performance of BvSP and the time consumed, the optimal choice is $k = 3$.

Threshold τ: This hyperparameter determines the confidence level of the quads in the final prediction $\mathcal{P}$. A higher threshold signifies that the quads predicted by the model are more frequently matched by multiple templates, thereby enhancing confidence in the correctness of the prediction of final quads. In the right plot of Figure 5, we observe that BvSP keeps increasing in accuracy and decreasing in recall as τ increases and produces optimal F1 score by maintaining within the range of 5 to 7. This indicates that setting τ to a larger value results in fewer screened quads but more evidence for each selected quad.

4.7 Evaluation on Other Datasets

We perform the ASQP task on the Rest15 and Rest16 datasets using baselines and the proposed BvSP in the full-shot setting. The results are presented in Table 4. BvSP continues to outperform all baselines in the full-shot setting. It is worth noting that BvSP exhibits superiority over MvP in both the top-3 and top-15. This underscores that BvSP not only excels in few-shot scenarios but also proves beneficial in full-shot scenarios.

Baseline	Rest15			Rest16		
	Pre	Rec	F1	Pre	Rec	F1
GAS	45.31	46.70	45.98	54.54	57.62	56.04
Paraphrase	46.16	47.72	46.93	56.63	59.30	57.93
DLO	47.08	49.33	48.18	57.92	**61.80**	59.79
ILO	47.78	**50.38**	49.05	57.58	61.17	59.32
MvP (top-3)	49.59	48.93	49.32	60.27	58.94	59.62
MvP (top-15)	-	-	51.04	-	-	60.39
BvSP (top-3)	54.63	47.53	50.83	63.59	59.35	61.40
BvSP (top-15)	**60.96**	47.15	**53.17**	**68.16**	59.42	**63.49**

Table 4: Full-shot results of ASQP task in the datasets Rest15 and Rest16.

Inputs-1		*These soups were flavorful and delish !*
Label-1		*(soups, flavorful, food_soup, great),* *(soups, delish, food_soup, great)*
Pred-1		*(soups, flavorful, food_soup, great),* *(soups, delish, food_soup, great)*
Target-1	T_1	*food_soup is great because soup is flavorful* [SSEP] *food_soup is great because soups is delish*
	T_2	[AT] *soups* [OT] *were flavorful* [AC] *food_soup* [SP] *great* [SSEP] [AT] *soups* [OT] *delish* [AC] *food_soup* [SP] *great*
	T_3	*(soups, flavorful, food_soup, great)* [SSEP] *(soups, delish, food_soup, great)*
Prediction-1		*(soups, flavorful, food_soup, great)*×2 *(soups, delish, food_soup, great)*×2 } Select *(soups, were flavorful, food_soup, great)*×1 *(soup, flavorful, food_bread, great)*×1 } Drop
Inputs-2		*it 's a beautiful building .*
Label-2		*(building, beautiful, building, great)*
Pred-2		*(room_overall, beautiful, building, great)*
Target-2	T_1	*room_overall is great because building is beautiful*
	T_2	[AT] *building* [OT] *beautiful* [AC] *room_overall* [SP] *great*
	T_3	*(building, beautiful, building, great)*
Prediction-2		*(building, beautiful, room_overall, great)*×2 } Select *(building, beautiful, building, great)*×1 } Drop

Figure 6: Two cases predicted by BvSP from the testing set of FSQP dataset under the one-shot settings.

4.8 Case Study

To completely understand the strengths and weaknesses of the BvSP, we conduct a case study. We provide one instance that demonstrates a successful prediction, while another instance showcases an error prediction by our method. These two cases are presented in Figure 6.

In case 1, we can find that two of the three templates have prediction errors (marked in red color). However, after passing our voting mechanism, the final prediction is corrected. This indicates clearly that BvSP shows a remarkable ability to successfully aggregate exactly predicted quads from multiple templates while simultaneously filtering out the error predictions among them by considering the commonalities between templates. Besides, in case 2, two out of the three templates are erroneously predicted (marked in red color), subsequently leading to an error final prediction. This case signifies that the accuracy of our methodology remains dependent on the rationality of a template. If multiple templates contain the same kind of error, this is hardly eliminated error by the voting mechanism. Thus, the observation highlights the fact that our aggregation strategy still possesses untapped potential for enhancement.

5 Related Work

5.1 Datasets for ASQP

The current ASQP datasets (Cai et al., 2021; Zhang et al., 2021a) are generated through similar processes and have a shared origin. Specifically, they are derived from SemEval challenge datasets (Pontiki et al., 2014, 2015, 2016), which provide aspect terms and corresponding sentiments for restaurant and laptop reviews. Fan et al. (2019) annotate the aspect-opinion pairs based on the SemEval datasets, and Peng et al. (2020) further expand sentiment tuples (at, ot, sp) based on (Fan et al., 2019). Zhang et al. (2021a) annotate Rest15 and Rest16 in light of the Semval tasks (Pontiki et al., 2015, 2016). Meanwhile, Cai et al. (2021) propose Restaurant and Laptop. Restaurant is based on initial SemEval challenge datasets and its extensions (Fan et al., 2019; Xu et al., 2020), while Cai et al. (2021) annotate the Laptop dataset based on the Amazon 2017 and 2018 reviews.

However, all of these datasets still contain many limitations, as explained in §2.3. The shortcomings of existing benchmark datasets motivate us to provide a more diverse set of reviews covering more categories and instances of each category.

5.2 Methods for ASQP

Aspect Sentiment Quad Predicition (ASQP) has received wide attention in recent years, learning four elements simultaneously, i.e. aspect sentiment quads. Corresponding solutions for ASQP can be divided into two categories: non-generation (Cai et al., 2021) and generation (Zhang et al., 2021a; Mao et al., 2022; Bao et al., 2022; Peper and Wang, 2022; Hu et al., 2022, 2023; Gou et al., 2023). The non-generation method employs two-stage framework to extract aspect sentiment quads by improving tradition-based methods such as Double-Propagation (Qiu et al., 2011b), JET (Xu et al., 2020), HGCN (Zhang et al., 2021a), TAS (Wan et al., 2020). Due to the simplicity and end-to-end manner, generation-based methods have become the main research direction. Promising works design novel approaches based on tree structure (Mao et al., 2022; Bao et al., 2022), contrastive learning (Peper and Wang, 2022), data augmentation (Hu et al., 2022), multi-view (Gou et al., 2023) and uncertainty (Hu et al., 2023). In contrast to these works, we approach the few-shot ASQP task from a broader perspective, introducing Broad-View Soft Prompting (BvSP) into our design.

6 Conclusion

This work studies the ASQP task from the few-shot perspective, which aims to handle unseen aspects with only a few supported samples. Therefore, we first build a new dataset called `FSQP`, which is specifically annotated for few-shot ASQP. Unlike existing ASQP datasets, `FSQP` provides a more balanced representation and covers a wider range of categories, thereby serving as a comprehensive benchmark for evaluating few-shot ASQP. Moreover, the generation-based paradigm has become the state-of-the-art technique for ASQP. However, previous methods overlook the correlation between different templates. In this study, we propose a broad-view soft prompting (BvSP) method to address this limitation. BvSP leverages the JS divergence to analyze the correlation among templates and selects relevant ones. It then guides the pre-trained language model with soft prompts based on these selected templates. Finally, the results are aggregated through voting. Extensive experiments conducted under few-shot settings demonstrate that BvSP exhibits universal effectiveness and substantial improvements in both explicit and implicit information.

Limitations

Our work represents the pioneering effort in tackling the few-shot ASQP task through the creation of a novel benchmark dataset, `FSQP`, and the introduction of a novel method, BvSP. However, it's essential to acknowledge that our work still has limitations, which can serve as valuable pointers for future research directions.

Firstly, we employ JS divergence to analyze the correlation between different templates. A smaller JS value signifies a stronger correlation between the two templates. Nonetheless, there may exist alternative criteria for template selection that could further boost the pre-trained language model, enhancing its support for the few-shot ASQP task.

Secondly, the experiments focus solely on the few-shot ASQP. However, compound ABSA includes numerous subtasks, such as aspect category sentiment analysis (Schmitt et al., 2018) and target aspect sentiment detection (Wan et al., 2020). Nevertheless, the `FSQP` dataset still satisfies the requirements for these tasks. Therefore, future research may consider exploring few-shot learning techniques for these tasks.

Lastly, it should be noted that utilizing multi-soft prompting introduces additional training and inference overheads, which scale proportionally with the number of selected templates. Despite this, our method relies solely on automatic optimization and does not raise human labor.

Acknowledgements

We sincerely thank all the anonymous reviewers for providing valuable feedback. This work is supported by the youth program of National Science Fund of Tianjin, China (Grant No. 22JC-QNJC01340).

References

Xiaoyi Bao, Z Wang, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 4044–4050.

Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 717–725.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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 (ACL-IJCNLP), pages 340–350.

Zhifang Fan, Zhen Wu, Xinyu 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.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. *arXiv preprint arXiv:2305.12627*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4803–4809. Association for Computational Linguistics.

Mengting Hu, Yinhao Bai, Yike Wu, Zhen Zhang, Liqi Zhang, Hang Gao, Shiwan Zhao, and Minlie Huang. 2023. Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction. *arXiv preprint arXiv:2306.00418*.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 7889–7900.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jianwei Li, Xianyong Li, Yajun Du, Yongquan Fan, Xiaoliang Chen, and Dong Huang. 2023. An aspect-category-opinion-sentiment quadruple extraction with distance information for implicit sentiment analysis. *Information Technology and Control*, 52(2):445–456.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Zheng Li, Ying Wei, Yu Zhang, Xiang Zhang, and Xin Li. 2019. Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4253–4260.

Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

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)*, pages 2215–2225.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8600–8607.

Joseph J Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. *arXiv preprint arXiv:2211.07743*.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation (SemEval 2016)*, pages 19–30.

Maria Pontiki, Dimitrios Galanis, Harris 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.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh. 2014. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pages 28–37.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011a. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011b. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, 37(1):9–27.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67.

Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. 2018. Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1114, Brussels, Belgium. Association for Computational Linguistics.

Thanveer Shaik, Xiaohui Tao, Christopher Dann, Haoran Xie, Yan Li, and Linda Galligan. 2022. Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, page 100003.

G Vinodhini and RM Chandrasekaran. 2012. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 9122–9129.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2023. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*.

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

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 (EMNLP)*, pages 9209–9219.

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 (ACL-IJCNLP)*, pages 504–510.

Tao Zhou and Kris MY Law. 2022. Semantic relatedness enhanced graph network for aspect category sentiment analysis. *Expert Systems with Applications*, 195:116560.

Datasets	#S	#Q	#C
Train	8204	11363	45
Dev	2204	2600	15
Test	2143	2420	20

Table 5: Data statistics. #S, #Q, and #C denote the number of sentences, quads, and aspect categories respectively.

Inputs Sentence	*The food is good.*
Quadruplet (*at, ot, ac, sp*)	(*food*, *good*, *food quality*, *positive*)
Semantic Mapping (x_{at}, x_{ot}, x_{ac}, x_{sp})	(*food*, *good*, *food quality*, *great*)
GAS	(*at, ot, ac, sp*)
Target sequence	(*food*, *good*, *food quality*, *positive*)
Paraphrase	x_{ac} *is* x_{sp} *because* x_{at} *is* x_{ot}
Target sequence	*food quality is great because food is good*
Special Symbols	[AT] x_{at} [OT] x_{ot} [AC] x_{ac} [SP] x_{sp}
Target sequence	[AT] *food* [OT] *good* [AC] *food quality* [SP] *great*

Figure 7: Template details of various methods.

A Datasets statistics

The original dataset FSQP is first divided into a training set, a development set, and a testing set according to aspect category. The statistics are displayed in Table 5. Few-shot learning aims to replicate real-world scenarios where the model encounters novel classes that were not included in the training dataset. Therefore, it can be observed that three sets have 45, 15, and 20 aspect categories without overlapping.

B Template details of various methods

The template details of each baseline method in Figure 7. It is worth noting that ILO and DLO also follow the special symbols templates but combine multiple template orders as data augmentation. For BvSP, we consider all possible templates, including one for GAS (Zhang et al., 2021b), one for Paraphrase (Zhang et al., 2021a), and twenty-four for Special Symbols templates (Hu et al., 2022) including all possible permutations, thus twenty-six in total.

C Supplementary Experiments

C.1 Analysis at Element-Level

In this section, the performance of BvSP is fully analyzed at four aspect-level elements: aspect term, opinion term, aspect category, and sentiment polarity. The corresponding analysis results are presented in Table 6.

Element	Method	One-shot	Two-shot	Five-shot	Ten-shot	Average
Aspect Term	GAS	79.44	81.64	84.04	85.42	82.63
	Para	79.64	82.73	85.41	86.77	83.63
	DLO	71.33	80.32	85.32	87.40	81.09
	ILO	**80.18**	82.45	84.36	86.64	83.40
	BvSP	79.41	**82.66**	**85.40**	**88.15**	**83.90**
Opinion Term	GAS	72.73	73.50	74.55	74.97	73.93
	Para	72.41	73.69	74.47	74.88	73.86
	DLO	64.99	70.39	73.96	74.66	71.00
	ILO	73.35	**74.11**	74.80	**75.20**	**74.36**
	BvSP	**73.63**	73.62	**74.85**	74.75	74.21
Aspect Category	GAS	43.99	69.17	83.75	86.60	70.87
	Para	38.88	65.65	82.85	86.37	68.43
	DLO	46.11	64.94	85.34	87.60	70.99
	ILO	38.90	60.99	83.05	86.98	67.48
	BvSP	**62.00**	**80.08**	**86.97**	**87.93**	**79.25**
Sentiment Polarity	GAS	80.63	81.71	81.76	82.27	81.59
	Para	80.63	81.21	81.18	82.06	81.27
	DLO	73.06	78.63	81.21	81.99	78.72
	ILO	**81.69**	81.70	81.76	82.33	81.87
	BvSP	80.71	**82.05**	**82.74**	**84.48**	**82.50**

Table 6: Analyses at aspect-based elements for few-shot ASQP in terms of F1 score (%).

Considering the aspect term and sentiment polarity first, BvSP shows superior performance compared to the baselines on the average F1 score. In addition, BvSP consistently surpasses the baselines in four few-shot scenarios under the aspect category. However, for the opinion term, our method only slightly less-performs the strong baseline ILO on the average F1 score. These indicate the effectiveness and robustness of BvSP on four elements from the unseen category.

Moreover, it is noticed that all methods achieve better performance in predicting aspect term, opinion term, and sentiment polarity in the one-shot scenario (more than 70% F1 scores), but they have poor performance for the aspect category (only around 40% F1 scores). This finding indicates that even in unseen aspect categories scenarios, the model can still effectively reason about the aspect term, the opinion term, and the sentiment polarity based on the knowledge and patterns acquired during the pre-training phase. A possible reason is that sentences share similar grammar rules and syntactic structures. Yet comprehending the semantics of unseen categories is struggling.

Finally, comparing from 1 to 10 shots, the performance improvement in the aspect category is more significant than the other three elements. This possibly shows that feeding neural networks the novel knowledge contributes more to their fast adaptation, pointing out the potential future direction.

C.2 Vote Mechanism

To investigate the effect of the aggregation method, we conducted further comparative experiments on

Method	One-shot	Two-shot	Five-shot	Ten-shot
BvSP	**37.99**	**47.34**	**53.52**	**54.43**
DLO (one path)	26.23	37.72	50.38	51.62
DLO (vote)	27.91	38.60	50.81	51.79

Table 7: Evaluation results (F1, %) of vote mechanism.

	Method	One-shot	Two-shot	Five-shot	Ten-shot
Explicit	GAS	26.74	41.13	50.55	52.67
	Para	24.16	39.40	51.19	53.69
	DLO	26.33	38.57	51.44	52.82
	ILO	25.41	37.53	50.74	53.09
	BvSP	**41.66**	**49.73**	**56.22**	**56.77**
Implicit	GAS	23.01	37.13	46.32	48.66
	Para	21.45	37.59	44.71	48.63
	DLO	25.72	35.43	47.44	48.52
	ILO	20.32	33.16	46.38	48.93
	BvSP	**31.53**	**49.05**	**48.45**	**51.44**

Table 8: Evaluation results on the explicit and implicit subsets in terms of F1 score (%).

the voting mechanism, comparing BvSP with the optimal baseline, DLO, which also utilizes multiple templates. The results of these comparisons are presented in Table 7. We can see that DLO (vote) performs better than DLO (one path), suggesting that employing the voting mechanism instead of choosing one path can enhance DLO's performance. However, while DLO exhibits some performance improvement, it still doesn't surpass our proposed BvSP method, further underscoring the effectiveness of our approach.

C.3 Implicit Information Prediction

To better understand the effectiveness of our proposed BvSP in predicting implicit information, it is separately evaluated in four different few-shot settings. The results of these comparisons are depicted in Table 8. Our testing set is divided into two subsets. The explicit subset refers to both the aspect term and opinion term being explicit. Then if any of them is implicit, it is put into the implicit subset.

We can observe that our method BvSP outperforms baselines under all four few-shot settings, demonstrating superior performance in both explicit and implicit information prediction. In particular, compared to the robust baseline GAS, BvSP achieves a notable enhancement of +8.32% and +6.34% in the average F1 score for explicit and implicit information, respectively.

Baseline	One-shot	Five-shot
ChatGPT	23.56	38.97
Llama-2-70B-Chat	22.91	25.95
Mixtral-8x7B-Instruct	17.38	23.15
GPT-4	28.50	40.09
BvSP-T5-Base	**37.99**	**53.52**
BvSP-Shearing-Llama	25.19	47.86

Table 9: Performance comparison of various language models under one-shot and five-shot settings. F1-score (%) metrics are reported for each model.

Furthermore, it's worth noting that BvSP's performance improves for both explicit and implicit subsets as the number of shots increases. The F1 score for the implicit subset shows a similar upward trend as for the explicit one. This finding highlights the model's reliance on ample data to enhance its performance. Otherwise, the uneven distribution of explicit and implicit information across quad categories in the FSQP dataset could explain the variations in results between models.

C.4 Performance of Different Language Models

We expand our evaluation to include three additional large language models: Llama-2-70B-Chat, Mixtral-8x7B-Instruct, and GPT-4. We also introduced a trainable LLM, shearing-llama-370M (Xia et al., 2023).

As shown in Table 9, BvSP-T5-Base consistently outperforms other models in both settings, achieving the highest F1 of 37.99% in the one-shot and 53.52% in the five-shot scenarios. This indicates its superiority to generalize from limited examples.

GPT-4 also shows strong performance, particularly in the five-shot setting, highlighting its robustness with additional context. Other models like ChatGPT and Llama-2-70B-Chat demonstrate moderate performance, with notable improvements in the five-shot setting compared to the one-shot. Mixtral-8x7B-Instruct performs lower overall. However, due to the inference length limitation of LLMs, errors often occur when the validated shot number exceeds five. Thus, we speculate that without additional training, increasing the inference length of LLMs could accommodate more shots and enhance the performance of LLM evaluations.

We also observe that the performance of BvSP-Shearing-Llama is not as good as BvSP-T5-Base.

Models	One-shot
GAS	216s
Paraphrase	201s
DLO	418s
ILO	436s
BvSP	477s

Table 10: Average running time of each model.

Two possible explanations are: (1) the effect of training parameters, suggesting that we might need to adjust them for Shearing-Llama, and (2) differences in model characteristics. Shearing-Llama is derived from the general model Llama-7b through pruning, focusing more on general tasks like lm-evaluation-harness. In contrast, T5-Base is pre-trained to solve downstream tasks, including sentiment analysis, giving it a potential advantage over Shearing-Llama.

C.5 Training Time Analysis

The average running time of each model is shown in Table 10. We can observe that on five generation-based methods, BvSP consistently causes more training time and consumes a little more time compared with ILO.

D Supplementary Materials of `FSQP`

For our dataset `FSQP`, we present the complete set of categories in the training set, development set, and testing set in Table 11. We further select two comments from each of the ten categories from these sets. The sampled reviews are demonstrated in Table 12, 13, and 14. The first column displays the review sentence. And the second column shows the extracted quads (aspect terms, aspect category, sentiment polarity, opinion terms).

	Aspect category
Train	room_overall, room_smoke, service_staff, service, price, salon, food_dessert, food, experience, hotel, service_staff_doctor, building_hall, food_meat_chicken, procedure_beauty_nails, food_meat, building_elevator, location, decor, room_bathroom, room_equipment, room_interior, procedure_relax_massage, procedure_beauty_nails_pedi, procedure_beauty_face, food_salad, service_staff_front-desk, food_bread, procedure_beauty_nails_mani, procedure_beauty_hair, food_meat_beef, drinks, parking, service_staff_master, procedure_beauty_wax, service_staff_owner, food_meat_pork, food_mealtype_start, cleanliness, food_meat_rib, food_portion, salon_equipment, room_bed, room_bedroom, food_fruit, internet
Dev	food_side_potato, restaurant, food_meat_steak, food_side_vegetables, food_meat_burger, food_cheese, food_side_pasta, food_mealtype_main, procedure_relax_train, food_sushi, salon_additional, food_seafood, salon_interior, salon_interior_bath, salon_interior_room
Test	food_eggs, food_soup, procedure_beauty_barber, service_management, procedure_relax_spa, salon_atmosphere, food_mealtype_breakfast, drinks_alcohol_light, entertainment_atmosphere, sport_pool, restaurant_atmosphere, food_selection, building, occasion, drinks_alcohol_beer, service_staff_waiter, drinks_non-alcohol, food_mealtype_dinner, drinks_alcohol_wine, service_wait

Table 11: The aspect categories contained in the training set, development set, and testing set are shown.

room_overall	
If you are the type to stay in your room a lot then this is the place for you .	(room, room_overall, positive, NULL)
We were very disappointed with our three day stay at red rock in room 16143 .	(room, room_overall, negative, very disappointed)
service	
They are very professional and great customer service .	(customer service, service, positive, professional and great)
We would have enjoyed more if not for the unprofessional wait staff .	(staff, service, negative, unprofessional)
price	
The prices are also reasonable .	(prices, price, positive, reasonable)
I did n't bother going back it seems like a waste of time and money ...	(money, price, negative, waste)
food	
I had the vodka penne and it was delicious !	(vodka penne, food, positive, delicious)
I have never had such disgusting chinese food then this .	(chinese food, food, negative, disgusting)
decor	
I 'm a huge phillip stark fan and i thought the decor was beautiful .	(decor, decor, positive, beautiful)
Currently staying here , carpet is gross .	(carpet, decor, negative, gross)
internet	
Wifi works best on odd room numbers .	(wifi, internet, positive, works best)
The sign on the wall gives you a wifi password to use , but it doesn't work .	(wifi password, internet, negative, doesn't work)
cleanliness	
First the cleanliness of the room was substandard .	(cleanliness, cleanliness, negative, substandard)
I was disappointed by the cleanliness of the room .	(cleanliness, cleanliness, negative, disappointed)
drinks	
The drinks were also delicious !	(drinks, drinks, positive, delicious)
Mmmm ... the drinks are n't that good!	(drinks, drinks, negative, n't that good)
location	
This is one of my favorite sephora locations .	(sephora locations, location, positive, favorite)
The only negative about this place is the location .	(location, location, negative, negative)
hotel	
We loved this hotel !	(hotel, hotel, positive, loved)
I will not be staying in this hotel again ...	(hotel, hotel, negative, not be staying in)

Table 12: Sampled reviews from the training set of FSQP.

food_side_potato	
We also ordered potatoes au gratin which were amazing ! ! !	(potatoes au gratin, food_side_potato, positive, amazing)
The potato skins were tiny and limp .	(potato skins, food_side_potato, negative, tiny and limp)
restaurant	
Acoustically , the restaurant was distracting .	(restaurant, restaurant, negative, distracting)
Finally a reliable chinese restaurant !	(chinese restaurant, restaurant, positive, reliable)
food_meat_steak	
The cut they suggested , was along the best steak i 've ever had .	(steak, food_meat_steak, positive, best)
The steak was so rough and disgusting i actually cut it up and fed it to my steaks .	(steak, food_meat_steak, negative, rough and disgusting)
food_side_vegetables	
The veggies with the sauces excellent !	(veggies, food_side_vegetables, positive, excellent)
I was a bit disappointed by the fried pickles	(fried pickles, food_side_vegetables, negative, disappointed)
food_meat_burger	
I ordered my wineburger it did not disappoint .	(wineburger, food_meat_burger, positive, not disappoint)
Burgers are very under cooked !	(burgers, food_meat_burger, negative, very under cooked)
food_cheese	
The cotija cheese was yummy .	(cotija cheese, food_cheese, positive, yummy)
The cheese had no flavor .	(cheese, food_cheese, negative, no flavor)
procedure_relax_train	
This gym has everything i need ... if only i could step it up so that i actually see results !	(gym, procedure_relax_train, positive, NULL)
The gym ... ... was crowded .	(gym, procedure_relax_train, negative, crowded)
food_sushi	
The sushi was really good !	(sushi, food_sushi, positive, good)
Cons - sushi pieces were smaller than i expected .	(sushi pieces, food_sushi, negative, smaller)
food_seafood	
The seafood fradiavolo was delicious .	(seafood fradiavolo, food_seafood, positive, delicious)
If you are expecting legal seafood do n't go here .	(seafood, food_seafood, negative, NULL)
salon_interior	
The bed is super comfortable .	(bed, salon_interior, positive, comfortable)
Seriously , it was the worst sofa bed in the world .	(sofa bed, salon_interior, negative, worst)

Table 13: Sampled reviews from the development set of FSQP.

food_eggs	
The eggs benedict were delicious !	(eggs benedict, food_eggs, positive, delicious)
The egg rolls were slighty burnt .	(egg rolls, food_eggs, negative, slighty burnt)
food_soup	
We had miso soup to start , was great !	(miso soup, food_soup, positive, great)
Avoid the french onion soup .	(french onion soup, food_soup, negative, avoid)
service_management	
Manager came by a few times to be sure we were satisfied .	(manager, service_management, positive, NULL)
Manager went out of her way to apologize but by then we were very very disappointed .	(manager, service_management, negative, disappointed)
procedure_relax_spa	
My friend and i had a perfect spa day package there .	(spa day, procedure_relax_spa, positive, perfect)
Do n't waste your time at any other spa on the strip .	(spa, procedure_relax_spa, negative, NULL)
drinks_alcohol_light	
Their signature pineapple martini was to die for !	(pineapple martini, drinks_alcohol_light, positive, to die for)
Martini was short and missing several sips .	(martini, drinks_alcohol_light, negative, short)
entertainment_atmosphere	
Also the noise level was a little high .	(noise level, entertainment_atmosphere, negative, little high)
Very nice people and great atmosphere !	(atmosphere, entertainment_atmosphere, positive, great)
sport_pool	
Pools are great in the summer .	(pools, sport_pool, positive, great)
The pool had trash in it .	(pool, sport_pool, negative, had trash in)
building	
It looks like an old mayan temple building in need of updates .	(mayan temple building, building, negative, old)
But it is a beautiful building .	(building, building, positive, beautiful)
food_selection	
Anything you want thats on the menu .	(menu, food_selection, positive, NULL)
But the menu is really very limited .	(menu, food_selection, negative, limited)
service_wait	
While you are waiting you can sit comfortably on one of the couches .	(waiting, service_wait, positive, comfortably)
So there i sat waiting even longer ! ! !	(waiting, service_wait, negative, even longer)

Table 14: Sampled reviews from the testing set of FSQP.