A Unified One-Step Solution for Aspect Sentiment Quad Prediction

Junxian Zhou[1], Haiqin Yang[2]*, Yuxuan He[1], Hao Mou[1], Junbo Yang[1]
[1]DataStory, Guangzhou, China
[2]International Digital Economy Academy, Shenzhen, China
{junius, heyuxuan, mouhao, junbo}@datastory.com.cn,
hqyang@ieee.org

Abstract

Aspect sentiment quad prediction (ASQP) is a challenging yet significant subtask in aspect-based sentiment analysis as it provides a complete aspect-level sentiment structure. However, existing ASQP datasets are usually small and low-density, hindering technical advancement. To expand the capacity, in this paper, we release two new datasets for ASQP, which contain the following characteristics: larger size, more words per sample, and higher density. With such datasets, we unveil the shortcomings of existing strong ASQP baselines and therefore propose a unified one-step solution for ASQP, namely One-ASQP, to detect the aspect categories and to identify the aspect-opinion-sentiment (AOS) triplets simultaneously. Our One-ASQP holds several unique advantages: (1) by separating ASQP into two subtasks and solving them independently and simultaneously, we can avoid error propagation in pipeline-based methods and overcome slow training and inference in generation-based methods; (2) by introducing sentiment-specific horns tagging schema in a token-pair-based two-dimensional matrix, we can exploit deeper interactions between sentiment elements and efficiently decode the AOS triplets; (3) we design "[NULL]" token can help us effectively identify the implicit aspects or opinions. Experiments on two benchmark datasets and our released two datasets demonstrate the advantages of our One-ASQP. The two new datasets are publicly released at https://www.github.com/Datastory-CN/ASQP-Datasets.

1 Introduction

Aspect-based sentiment analysis (ABSA) is a critical fine-grained opinion mining or sentiment analysis problem that aims to analyze and understand people's opinions or sentiments at the aspect level (Liu, 2012; Pontiki et al., 2014; Zhang et al., 2022). Typically, there are four fundamental sentiment elements in ABSA: (1) *aspect category* (c) defines the type of the concerned aspect; (2) *aspect term* (a) denotes the opinion target which is explicitly or implicitly mentioned in the given text; (3) *opinion term* (o) describes the sentiment towards the aspect; and (4) *sentiment polarity* (s) depicts the sentiment orientation. For example, given an opinionated sentence, "touch screen is not sensitive," we can obtain its (c, a, o, s)-quadruple as ("Screen#Sensitivity", "touch screen", "not sensitive", NEG), where NEG indicates the negative sentiment polarity.

Task	Output	Example Output
ATE	$\{a\}$	{touch screen}
ACD	$\{c\}$	{Screen#Sensitivity}
AOPE	$\{(a, o)\}$	{(touch screen, not sensitive)}
ACSA	$\{(c, s)\}$	{(Screen#Sensitivity, NEG)}
E2E-ABSA	$\{(a, s)\}$	{(touch screen, NEG)}
ASTE	$\{(a, o, s)\}$	{(touch screen, not sensitive, NEG)}
TASD	$\{(c, a, s)\}$	{(Screen#Sensitivity, touch screen, NEG)}
ASQP/ ACOS	$\{(c, a, o, s)\}$	{(Screen#Sensitivity, touch screen, not sensitive, NEG)}

Table 1: The outputs of an example, "touch screen is not sensitive", for various ABSA tasks. a, c, o, s, and NEG are defined in the first paragraph of Sec. 1.

Due to the rich usage of applications, numerous research efforts have been made on ABSA to predict or extract fine-grained sentiment elements (Jiao et al., 2019; Pontiki et al., 2014, 2015, 2016; Zhang et al., 2022; Yang et al., 2021). Based on the number of sentimental elements to be extracted, existing studies can be categorized into the following tasks: (1) *single term extraction* includes aspect term extraction (ATE) (Li and Lam, 2017; He et al., 2017), aspect category detection (ACD) (He et al., 2017; Liu et al., 2021); (2) *pair extraction* includes aspect-opinion pairwise extraction (AOPE) (Yu et al., 2019; Wu et al., 2020), aspect-category sentiment analysis (ACSA) (Cai et al., 2020; Dai et al., 2020), and

*The corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 12249–12265

End-to-End ABSA (E2E-ABSA) (Li et al., 2019b; He et al., 2019) to extract the aspect and its sentiment; (3) *triplet extraction* includes aspect-sentiment triplet extraction (ASTE) (Mukherjee et al., 2021; Chen et al., 2021), and Target Aspect Sentiment Detection (TASD) (Wan et al., 2020); (4) *quadruple extraction* includes aspect-category-opinion-sentiment (ACOS) quadruple extraction (Cai et al., 2021) and aspect sentiment quad prediction (ASQP) (Zhang et al., 2021a). ACOS and ASQP are the same tasks, which aim to extract all aspect-category-opinion-sentiment quadruples per sample. Since ASQP covers the whole task name, we use ASQP to denote the ABSA quadruple extraction task. Table 1 summarizes an example of the outputs of various ABSA tasks.

This paper focuses on ASQP because it provides a complete aspect-level sentiment analysis (Zhang et al., 2022). We first observe that existing ASQP datasets are crawled from only one source and are small with low-density (Cai et al., 2021; Zhang et al., 2021a). For example, the maximum sample size is around 4,000, while the maximum number of quadruples per sample is around 1.6. This limits the technical development of ASQP. Second, ASQP includes two extraction subtasks (aspect extraction and opinion extraction) and two classification subtasks (category classification and sentiment classification). Modeling the four subtasks simultaneously is challenging, especially when the quadruples contain implicit aspects or opinions (Cai et al., 2021). Though existing studies can resolve ASQP via pipeline-based (Cai et al., 2021) or generation-based methods (Zhang et al., 2021a; Mao et al., 2022; Bao et al., 2022; Gao et al., 2022), they suffer from different shortcomings, i.e., pipeline-based methods tend to yield error propagation while generation-based methods perform slowly in training and inference.

To tackle the above challenges, we first construct two datasets, **en-Phone** and **zh-FoodBeverage**, to expand the capacity of datasets. en-Phone is an English ASQP dataset in the cell phone domain collected from several e-commercial platforms, while zh-FoodBeverage is the first Chinese ASQP dataset collected from multiple sources under the categories of Food and Beverage. Compared to the existing ASQP datasets, our datasets have 1.75 to 4.19 times more samples and a higher quadruple density of 1.3 to 1.8. This achievement is a result of our meticulous definition and adherence to annotation guidelines, which allow us to obtain more fine-grained quadruples.

After investigating strong ASQP baselines, we observed a decline in performance on our newly released dataset. This finding, coupled with the shortcomings of the existing baselines, motivated us to develop a novel one-step solution for ASQP, namely One-ASQP. As illustrated in Fig. 1, our One-ASQP adopts a shared encoder from a pre-trained language model (LM) and resolves two tasks, aspect category detection (ACD) and aspect-opinion-sentiment co-extraction (AOSC) simultaneously. ACD is implemented by a multi-class classifier and AOSC is fulfilled by a token-pair-based two-dimensional (2D) matrix with the sentiment-specific horns tagging schema, a popular technique borrowed from the joint entity and relation extraction (Wang et al., 2020; Shang et al., 2022). The two tasks are trained independently and simultaneously, allowing us to avoid error propagation and overcome slow training and inferring in generation-based methods. Moreover, we also design a unique token, "[NULL]", appending at the beginning of the input, which can help us to identify implicit aspects or opinions effectively.

Our contributions are three-fold: (1) We construct two new ASQP datasets consisting of more fine-grained samples with higher quadruple density while covering more domains and languages. Significantly, the released zh-FoodBeverage dataset is the first Chinese ASQP dataset, which provides opportunities to investigate potential technologies in a multi-lingual context for ASQP. (2) We propose One-ASQP to simultaneously detect aspect categories and co-extract aspect-opinion-sentiment triplets. One-ASQP can absorb deeper interactions between sentiment elements without error propagation and conquer slow performance in generation-based methods. Moreover, the delicately designed "[NULL]" token helps us to identify implicit aspects or opinions effectively. (3) We conducted extensive experiments demonstrating that One-ASQP is efficient and outperforms the state-of-the-art baselines in certain scenarios.

2 Datasets

We construct two datasets [1] to expand the capacity of existing ASQP datasets.

[1] More details are provided in Appendix A.

	#s	#w/s	#c	#q	#q/s	EA&EO	EA&IO	IA&EO	IA&IO	#NEG	#NEU	#POS	Avg. #w/a	Avg. #w/o
Restaurant-ACOS	2,286	15.11	13	3,661	1.60	2,431	350	530	350	1,007	151	2,503	1.46	1.20
Laptop-ACOS	4,076	15.73	121	5,773	1.42	3,278	1,241	912	342	1,879	316	3,578	1.40	1.09
en-Phone	7,115	25.78	88	15,884	2.23	13,160	2,724	-	-	3,751	571	11,562	1.73	1.98
zh-FoodBeverage	9,570	193.95	138	24,973	2.61	17,407	7,566	-	-	6,778	-	18,195	2.60	2.04

Table 2: Data statistics for the ASQP task. # denotes the number of corresponding elements. s, w, c, q stand for samples, words, categories, and quadruples, respectively. EA, EO, IA, and IO denote explicit aspect, explicit opinion, implicit aspect, and implicit opinion, respectively. "-" means this item is not included.

2.1 Source

en-Phone is an English dataset collected from reviews on multiple e-commerce platforms in July and August of 2021, covering 12 cell phone brands. To increase the complexity and the quadruple density of the dataset, we deliver the following filtering steps: (1) applying the LangID toolkit [2] to filter out comments whose body content is not in English; (2) filtering out samples with less than 8 valid tokens. **zh-FoodBeverage** is the first Chinese ASQP dataset, collected from Chinese comments in multiple sources in the years 2019-2021 under the categories of Food and Beverage. We clean the data by (1) filtering out samples with lengths less than 8 and greater than 4000; (2) filtering out the samples with valid Chinese characters less than 70%; (3) filtering out ad texts by a classifier which is trained by marketing texts with 90% classification accuracy.

2.2 Annotation

A team of professional labelers is asked to label the texts following the guidelines in Appendix A.2. Two annotators individually annotate the same sample by our internal labeled system. The strict quadruple matching F1 score between two annotators is 77.23%, which implies a substantial agreement between two annotators (Kim and Klinger, 2018). In case of disagreement, the project leader will be asked to make the final decision. Some typical examples are shown in Table 10.

2.3 Statistics and Analysis

Table 2 reports the statistics of two existing ASQP benchmark datasets and our released datasets for comparison. **en-Phone** contains 7,115 samples with 15,884 quadruples while **zh-FoodBeverage** contains 9,570 samples with 24,973 quadruples. The size and the number of quadruples are significantly larger than the current largest ASQP benchmark dataset, i.e., Laptop-ACOS. The statistics show that our released datasets contain unique characteristics and are denser than existing Restaurant-ACOS and Laptop-ACOS: (1) the number of words per sample is 25.78 and 193.95 for en-Phone and zh-FoodBeverage, respectively, while the number of quadruples per sample is 2.23 and 2.61 for en-Phone and zh-FoodBeverage accordingly. This shows that en-Phone and zh-FoodBeverage are much denser than existing ASQP datasets; (2) based on the annotation guidelines, we only label opinionated sentences with explicit aspects. Moreover, due to commercial considerations, we exclude sentences with neutral sentiment in zh-FoodBeverage; (3) here, we define more fine-grained aspects and opinions than existing ASQP datasets; see more examples in Appendix A. Consequently, we attain a longer average length per aspect and per opinion, as reported in the last two columns of Table 2.

3 Methodology

3.1 ASQP Formulation

Given an opinionated sentence $\mathbf{x}$, ASQP is to predict all aspect-level sentiment quadruples $\{(c, a, o, s)\}$, which corresponds to the aspect category, aspect term, opinion term, and sentiment polarity, respectively. The aspect category c belongs to a category set $\mathcal{C}$; the aspect term a and the opinion term o are typically text spans in $\mathbf{x}$ while they can be null if the target is not explicitly mentioned, i.e., $a \in V_{\mathbf{x}} \cup \{\emptyset\}$ and $o \in V_{\mathbf{x}} \cup \{\emptyset\}$, where $V_{\mathbf{x}}$ denotes the set containing all possible continuous spans of $\mathbf{x}$. The sentiment polarity s belongs to one of the sentiment classes, SENTIMENT={POS, NEU, NEG}, which corresponds to the positive, neutral, and negative sentiment, respectively.

3.2 One-ASQP

Our One-ASQP resolves two subtasks, ACD and AOSC, simultaneously, where ACD seeks a classifier to determine the aspect categories, and AOSC is to extract all (a, o, s)-triplets.

[2] https://pypi.org/project/langid/

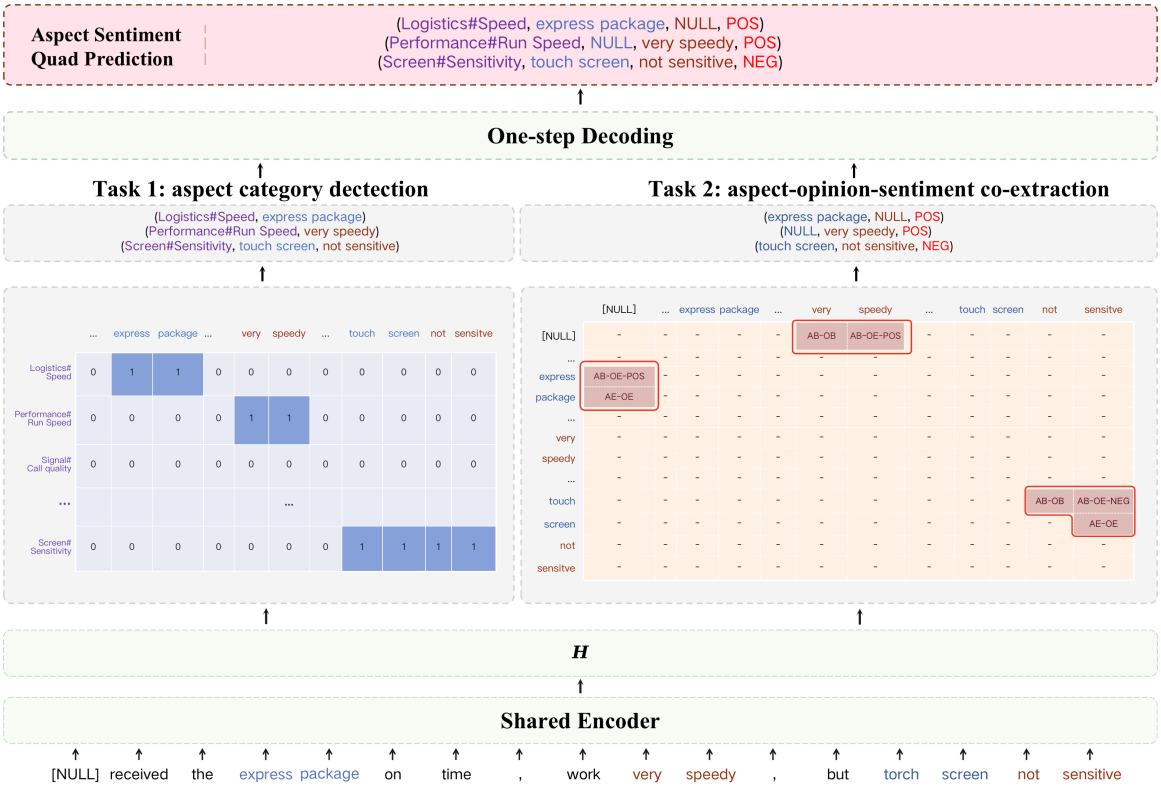


Figure 1: The structure of our One-ASQP: solving ACD and AOSC simultaneously. ACD is implemented by a multi-class classifier. AOSC is fulfilled by a token-pair-based 2D matrix with sentiment-specific horns tagging. The results in the row of "[NULL]" indicate no aspect for the opinion of "very speedy". In contrast, the results in the column of "[NULL]" imply no opinion for the aspect of "express package".

Given $\mathbf{x}$ with n-tokens, we construct the input as follows:

$$[\text{NULL}]\, x_1\, x_2\, \ldots\, x_n, \tag{1}$$

where the token [NULL] is introduced to detect implicit aspects or opinions; see more details in Sec. 3.2.2. Now, via a pre-trained LM, both tasks share a common encoder to get the representations:

$$\mathbf{H} = \mathbf{h}_{\text{NULL}}\, \mathbf{h}_1\, \mathbf{h}_2\, \ldots\, \mathbf{h}_n \in \mathbb{R}^{d\times(n+1)}, \tag{2}$$

where d is the token representation size.

3.2.1 Aspect Category Detection

We apply a classifier to predict the probability of category detection:

$$\mathbf{C} = \text{sigmoid}(\mathbf{W}_2(\text{RELU}(\mathbf{W}_1\mathbf{H} + \mathbf{b}_1))), \tag{3}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times d}$, $\mathbf{b}_1 \in \mathbb{R}^d$, $\mathbf{W}_2 \in \mathbb{R}^{|\mathcal{C}|\times d}$. Here, $|\mathcal{C}|$ is the number of categories in $\mathcal{C}$. Hence, $\mathbf{C} \in \mathbb{R}^{|\mathcal{C}|\times(n+1)}$, where C_{ij} indicates the probability of the i-th token to the j-th category.

3.2.2 AOSC

We tackle AOSC via a token-pair-based 2D matrix with the sentiment-specific horns tagging schema to determine the positions of aspect-opinion pairs and their sentiment polarity.

Tagging We define four types of tags: (1) **AB-OB** denotes the cell for the beginning position of an aspect-opinion pair. For example, as ("touch screen", "not sensitive") is an aspect-opinion pair, the cell corresponding to ("touch", "not") in the 2D matrix is marked by "AB-OB". (2) **AE-OE** indicates the cell for the end position of an aspect-opinion pair. Hence, the cell of ("screen", "sensitive") is marked by "AE-OE". (3) **AB-OE-*SENTIMENT** defines a cell with its sentiment polarity, where the row position denotes the beginning of an aspect and the column position denotes the end of an opinion. Hence, the cell of ("touch", "sensitive") is tagged by "AB-OE-NEG". As SENTIMENT consists of three types of sentiment polarity, there are three cases in AB-OE-*SENTIMENT.

(4) **"-"** denotes the cell other than the above three types. Hence, we have five types of unique tags, {AB-OB, AE-OE, AB-OE-POS, AB-OE-NEU, AB-OE-NEG}.

Triplet Decoding Since the tagged 2D matrix has marked the boundary tokens of all aspect-opinion pairs and their sentiment polarity, we can decode the triples easily. First, by scanning the 2D matrix column-by-column, we can determine the text spans of an aspect, starting with "AB-OE-*SENTIMENT" and ending with "AE-OE". Similarly, by scanning the 2D matrix row-by-row, we can get the text spans of an opinion, which start from "AB-OB" and end with "AB-OE-*SENTIMENT". Finally, the sentiment polarity can be easily determined by "AB-OE-*SENTIMENT".

Implicit Aspects/Opinions Extraction Detecting implicit aspects or opinions is critical in ASQP (Cai et al., 2021). Here, we append the "[NULL]" token at the beginning of the input. Our One-ASQP can then easily determine the cases of Implicit Aspects and Explicit Opinions (IA&EO) and Explicit Aspects and Implicit Opinions (EA&IO). The whole procedure is similar to the above triplet decoding: when the text spans at the row of "[NULL]" start from "AB-OB" and end with "AB-OE-*SENTIMENT", we can obtain an explicit opinion without aspect. Meanwhile, when the text spans at the column of "[NULL]" start from "AB-OE-*SENTIMENT" and ends with "AE-OE", we can obtain an explicit aspect without opinion. As shown in Fig. 1, we can quickly obtain the corresponding aspect-opinion pairs as "(NULL, very speedy)" and "(express package, NULL)". The sentiment polarity can also be determined by "AB-OE-*SENTIMENT" accordingly. Although the current setting for IA&IO cannot be solved directly, it is possible to resolve it in two steps. First, we can identify IA&IO using tools such as Extract-Classify-ACOS (Cai et al., 2021). Then, we can classify aspect categories and sentiment polarity. However, a unified solution with One-ASQP is left for future work.

Tagging Score Given $\mathbf{H}$, we compute the probabilities of the (i, j)-th cell to the corresponding tags by:

$$\mathbf{a}_i = \mathbf{W}_a \mathbf{h}_i + \mathbf{b}_a, \tag{4}$$

$$\mathbf{o}_j = \mathbf{W}_o \mathbf{h}_j + \mathbf{b}_o, \tag{5}$$

$$\mathbf{P}_{ij} = \text{sigmoid}(\mathbf{a}_i^T \mathbf{o}_j) \in \mathbb{R}^5 \tag{6}$$

where $\mathbf{W}_a \in \mathbb{R}^{D \times d}$ and $\mathbf{W}_o \in \mathbb{R}^{D \times d}$ are the weight matrices for the aspect token and the opinion token, respectively, $\mathbf{b}_a \in \mathbb{R}^D$ and $\mathbf{b}_o \in \mathbb{R}^D$ are the biases for the aspect token and the opinion token, respectively. D is the hidden variable size set to 400 as default.

3.3 Training Procedure

Training We train ACD and AOSC jointly by minimizing the following loss function:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{ACD} + \beta \mathcal{L}_{AOSC}, \tag{7}$$

where α and β are trade-off constants set to 1 for simplicity. The ACD loss $\mathcal{L}_{ACD}$ and the AOSC loss $\mathcal{L}_{AOSC}$ are two cross-entropy losses defined as follows:

$$\mathcal{L}_{ACD} = -\frac{1}{n \times |\mathcal{C}|} \times \sum_{i=1}^{n} \sum_{j=0}^{|\mathcal{C}|-1} \{ y_{ij}^{\mathcal{C}} \log C_{ij} + (1 - y_{ij}^{\mathcal{C}}) \log(1 - C_{ij}) \}, \tag{8}$$

$$\mathcal{L}_{AOSC} = -\frac{1}{(n+1) \times (n+1) \times 5} \times \sum_{i=0}^{n} \sum_{j=0}^{n} \{ \mathbf{Y}_{ij}^{t} \log \mathbf{P}_{ij} + (1 - \mathbf{Y}_{ij}^{t}) \log(1 - \mathbf{P}_{ij}) \}, \tag{9}$$

where C_{ij} is the predicted category computed by Eq. (3), $y_{ij}^{\mathcal{C}} \in \{0, 1\}$ and it is 1 when the i-th token is assigned to the j-th category and 0 otherwise. $\mathbf{P}_{ij}$ is the predicted tagging score computed by Eq. (6) for all five types of tags while $\mathbf{Y}_{ij} \in \mathbb{R}^5$ is the ground-truth one-hot encoding.

During training, we implement the negative sampling strategy as (Li et al., 2021) to improve the performance of our One-ASQP on unlabeled quadruples. We set the negative sampling rate to 0.4, a suggested range in (Li et al., 2021) that has yielded good results. Specifically, to minimize the loss in Eq.(7), we randomly sample 40% of unlabeled entries as negative instances, which correspond to '0' in ACD and '-' in AOSC, as shown in Fig.1.

3.4 Quadruples Decoding

After attaining the model, we can obtain the category sequences of ACD and the AOS triplets in the AOSC matrix simultaneously. We then decode the quadruples in one step via their common terms. For example, as shown in Fig. 1, we can merge (Logistics#Speed, express package) and (express package, NULL, POS) via the common aspect term,

"express package", and obtain the quadruple (Logistics#Speed, express package, NULL, POS).

Overall, our One-ASQP consists of two independent tasks, ACD and AOSC. Their outputs only share in the final decoding stage and do not rely on each other during training as the pipeline-based methods need. This allows us to train the model efficiently and decode the results consistently in both training and test.

4 Experimental Setup

Datasets We conduct the experiments on four datasets in Table 2. For Restaurant-ACOS and Laptop-ACOS, we apply the original splitting on the training, validation, and test sets (Cai et al., 2021). For en-Phone and zh-FoodBeverage, the splitting ratio is 7:1.5:1.5 for training, validation, and test, respectively.

Evaluation Metrics We employ F1 scores as the main evaluation metric and also report the corresponding **P**recision and **R**ecall scores. A sentiment quad prediction is counted as correct if and only if all the predicted elements are exactly the same as the gold labels. The time cost is also recorded to demonstrate the efficiency of One-ASQP.

Implementation Details One-ASQP is implemented by PyTorch 1.13.1. All experiments are run on a workstation with an Intel Xeon E5-2678 v3@2.50GHz CPU, 256G memory, a single A5000 GPU, and Ubuntu 20.04. For English datasets, we adopt LMs of DeBERTaV3-base and DeBERTaV3-large (He et al., 2021), which contain 12 layers with a hidden size of 768 and 24 layers with a hidden size of 1024, respectively. For the Chinese dataset, we adopt MacBERT (Cui et al., 2020), a Chinese LM with the same structure as DeBERTaV3. For the English datasets, the maximum token length is set to 128 as the maximum average word length is only 25.78, as shown in Table 2. For the zh-FoodBeverage, the maximum token length is 256. The batch size and learning rate for all experiments are [32, 3e-5] as they can perform well. We monitor the F1 score on the validation set and terminate the training when no score drops for four epochs. Finally, we report the scores on the test set by the best model on the validation set.

Baselines We compare our One-ASQP with strong baselines: (1) *pipeline-based methods* consist of four methods, i.e., **DP-ACOS**, **JET-ACOS**, **TAS-BERT-ACOS**, and **Extract-Classify-ACOS**, which are all proposed in (Cai et al., 2021); (2) *generation-based methods* include BART for ABSA (**BARTABSA**) (Yan et al., 2021), Generative Aspect-based Sentiment analysis (**GAS**) (Zhang et al., 2021b), **Paraphrase** generation for ASQP (Zhang et al., 2021a), **Seq2Path** (Mao et al., 2022), **GEN-SCL-NAT** (Peper and Wang, 2022), and ABSA with Opinion Tree Generation (**OTG**) (Bao et al., 2022).

Method	Restaurant-ACOS			Laptop-ACOS		
	P	R	F1	P	R	F1
DP-ACOS	34.67	15.08	21.04	13.04	0.57	8.00
JET-ACOS	59.81	28.94	39.01	44.52	16.25	23.81
TAS-BERT-ACOS	26.29	46.29	33.53	**47.15**	19.22	27.31
Extract-Classify-ACOS	38.54	52.96	44.61	45.56	29.48	35.80
BARTABSA	56.62	55.35	55.98	41.65	40.46	41.05
GAS	60.69	58.52	59.59	41.60	42.75	42.17
Paraphrase	58.98	59.11	59.04	41.77	**45.04**	43.34
Seq2Path	-	-	58.41	-	-	42.97
GEN-SCL-NAT	-	-	62.62	-	-	45.16
OTG	63.96	**61.74**	**62.83**	46.11	44.79	**45.44**
One-ASQP (base)	62.60	57.21	59.78	42.83	40.00	41.37
One-ASQP (large)	**65.91**	56.24	60.69	43.80	39.54	41.56

Table 3: Results of Restaurant-ACOS and Laptop-ACOS. Scores are averaged over 5 runs with different seeds.

5 Results and Discussions

5.1 Main Results

Method	en-Phone			zh-FoodBeverage		
	P	R	F1	P	R	F1
Extract-Classify-ACOS	31.28	33.23	32.23	41.39	32.53	36.43
Paraphrase	46.72	49.84	48.23	52.74	50.47	51.58
GEN-SCL-NAT	45.16	**51.56**	48.15	54.28	48.95	51.48
One-ASQP (base)	**57.90**	49.86	53.58	56.51	**59.13**	57.79
One-ASQP (large)	57.42	50.96	**54.00**	**60.96**	56.24	**58.51**

Table 4: Results of en-Phone and zh-FoodBeverage. Scores are averaged over five runs with different seeds.

Table 3 reports the comparison results on two existing ASQP datasets. Since all methods apply the same splitting on these two datasets, we copy the results of baselines from corresponding references. The results show that: (1) Generation-based methods gain significant improvement over pipeline-based methods as pipeline-based methods tend to propagate the errors. (2) Regarding generation-based methods, OTG attains the best performance on the F1 score. The exceptional performance may come from integrating various features, e.g., syntax and semantic information, for forming the opinion tree structure (Bao et al., 2022). (3) Our One-ASQP is competitive with generation-based methods. By checking the LM sizes, we know that the generation-based baselines except BARTABSA apply T5-base as the LM, which consists of 220M

parameters. In comparison, our One-ASQP model utilizes DeBERTaV3, which consists of only 86M and 304M backbone parameters for its base and large versions, respectively. The compact model parameter size is a crucial advantage of our approach. However, on the Restaurant-ACOS and Laptop-ACOS datasets, One-ASQP falls slightly behind some generation-based methods that can take advantage of the semantics of sentiment elements by generating natural language labels. In contrast, One-ASQP maps each label to a specific symbol, similar to the numerical indexing in classification models. Unfortunately, the limited quantity of these datasets prevents our One-ASQP model from achieving optimal performance.

We further conduct experiments on en-Phone and zh-FoodBeverage and compare our One-ASQP with three strong baselines, Extract-Classify-ACOS, Paraphrase, and GEN-SCL-NAT. We select them because Extract-Classify-ACOS is the best pipeline-based method. Furthermore, Paraphrase and GEN-SCL-NAT are two strong generation-based baselines releasing source codes, which is easier for replication. Results in Table 4 are averaged by five runs with different random seeds and show that our One-ASQP, even when adopting the base LM version, outperforms three strong baselines. We conjecture that good performance comes from two reasons: (1) The newly-released datasets contain higher quadruple density and fine-grained sentiment quadruples. This increases the task difficulty and amplifies the order issue in generation-based methods (Mao et al., 2022), i.e., the orders between the generated quads do not naturally exist, or the generation of the current quads should not condition on the previous ones. More evaluation tests are provided in Sec. 5.4. (2) The number of categories in the new datasets is much larger than Restaurant-ACOS and Laptop-ACOS. This also increases the search space, which tends to yield generation bias, i.e., the generated tokens neither come from the original text nor the pre-defined categories and sentiments. Overall, the results demonstrate the significance of our released datasets for further technical development.

Table 5 reports the time cost (in seconds) of training in one epoch and inferring the models on Restaurant-ACOS and en-Phone; more results are in Appendix B.1. The results show that our One-ASQP is much more efficient than the strong baselines as Extract-Classify-ACOS needs to encode twice and Paraphrase can only decode the token sequentially. To provide a fair comparison, we set the batch size to 1 and show the inference time in the round bracket. The overall results show that our One-ASQP is more efficient than the baselines. Our One-ASQP can infer the quadruples parallel, which is much favorite for real-world deployment.

Method	Restaurant-ACOS		en-Phone	
	Train	Inference	Train	Inference
Extract-Classify-ACOS	38.43	14.79	158.34	25.23
Paraphrase	30.52	58.23	99.23	160.56
GEN-SCL-NAT	35.32	61.64	104.23	175.23
OneASQP (base)	11.23	6.34 (29.35)	32.23	6.32 (35.45)
OneASQP (large)	17.63	14.63 (44.62)	105.23	10.34 (61.23)

Table 5: Time cost (seconds) on Restaurant-ACOS and en-Phone. For a fair comparison with baselines, we record the inference time of our One-ASQP with the batch size of 1 and report them in the round bracket.

5.2 Effect of Handling Implicit Aspects/Opinions

Table 6 reports the breakdown performance of the methods in addressing the implicit aspects/opinions problem. The results show that (1) the generation-based baseline, GEN-SCL-NAT, handles EA&IO better than our One-ASQP when the quadruple density is low. Accordingly, One-ASQP performs much better than GEN-SCL-NAT on IA&EO in Restaurant-ACOS. GEN-SCL-NA performs worse in IA&EO may be because the generated decoding space of explicit opinions is huge compared to explicit aspects. (2) In en-Phone and zh-FoodBeverage, One-ASQP consistently outperforms all baselines on EA&EO and EA&IO. Our One-ASQP is superior in handling implicit opinions when the datasets are more fine-grained.

5.3 Ablation Study on ADC and AOSC

To demonstrate the beneficial effect of sharing the encoder for ADC and AOS tasks. We train these two tasks separately, i.e., setting (α, β) in Eq. 7 to (1.0, 0.0) and (0.0, 1.0). The results in Table 7 show that our One-ASQP absorbs deeper information between two tasks and attains better performance. By sharing the encoder and conducting joint training, the connection between the category and other sentiment elements can become more tightly integrated, thereby contributing to each other.

5.4 Effect of Different Quadruple Densities

We conduct additional experiments to test the effect of different quadruple densities. Specifically, we keep those samples with only one quadruple

Method	Restaurant-ACOS			Laptop-ACOS			en-Phone		zh-FoodBeverage	
	EA&EO	EA&IO	IA&EO	EA&EO	EA&IO	IA&EO	EA&EO	EA&IO	EA&EO	EA&IO
Extract-Classify	45.0	23.9	34.7	35.4	16.8	39.0	35.2	24.2	37.2	33.3
Paraphrase	65.4	45.6	53.3	45.7	33.0	51.0	49.1	45.6	50.9	49.9
GEN-SCL-NAT	**66.5**	**46.2**	56.5	**45.8**	**34.3**	**54.0**	50.1	45.4	50.9	49.9
One-ASQP	66.3	31.1	**64.2**	44.4	26.7	53.5	**54.8**	**52.9**	**55.4**	**59.8**

Table 6: Breakdown performance (F1 scores) to depict the ability to handle implicit aspects or opinions. E and I stand for Explicit and Implicit, respectively, while A and O denote Aspect and Opinion, respectively.

	Restaurant-ACOS		Laptop-ACOS		en-Phone		zh-FoodBeverage	
	ADC F1	AOS F1	ADC F1	AOS F1	ADC F1	AOS F1	ADC F1	AOS F1
One-ASQP ($\alpha = 1.0, \beta = 0.0$)	68.64	-	47.45	-	63.43	-	64.57	-
One-ASQP ($\alpha = 0.0, \beta = 1.0$)	-	63.14	-	63.03	-	54.06	-	56.81
One-ASQP (base)	**75.85**	**65.88**	**51.62**	**65.13**	**66.09**	**57.99**	**66.90**	**62.62**

Table 7: Ablation study of One-ASQP on two losses.

in en-Phone and zh-FoodBeverage and construct two lower-density datasets, en-Phone (one) and zh-FoodBeverage (one). We then obtain 1,528 and 3,834 samples in these two datasets, respectively, which are around one-fifth and two-fifth of the original datasets accordingly.

We only report the results of our OneASQP with the base versions of the corresponding LMs and Paraphrase. Results in Table 8 show some notable observations: (1) Paraphrase can attain better performance on en-Phone (one) than our OneASQP. It seems that generation-based methods are powerful in the low-resource scenario. However, the performance is decayed in the full datasets due to the generation order issue. (2) Our One-ASQP significantly outperforms Paraphrase in zh-FoodBeverage for both cases. The results show that our OneASQP needs sufficient training samples to perform well. However, in zh-FoodBeverage (one), the number of labeled quadruples is 3,834. The effort is light in real-world applications.

Method	en-Phone		zh-FoodBeverage	
	one	full	one	full
Paraphrase	**49.78**	48.23	49.23	50.23
OneASQP	36.12	**53.58**	**53.39**	**57.79**

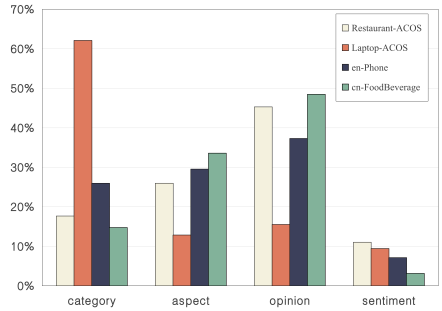
Table 8: Comparison results on different datasets with different quadruple densities.

5.5 Error Analysis and Case Study

To better understand the characteristics of our One-ASQP, especially when it fails. We conduct the error analysis and case study in this section. We check the incorrect quad predictions on all datasets and show one typical error example for each type from Laptop-ACOS in Fig. 2, where we report the percentage of errors for better illustration. The results show that (1) In general, extracting aspects and opinions tends to introduce larger errors than classifying categories and sentiments. Aspects and opinions have more complex semantic definitions than categories and sentiments, and extracting implicit cases further increases the difficulty of these tasks. (2) There is a significant category error in Laptop-ACOS, likely due to an imbalance issue in which there are 121 categories with relatively small samples per category. For example, 35 categories have less than two quadruples. (3) The percentage of opinion errors is higher than that of aspect errors in all datasets because opinions vary more than aspects, and there are implicit opinions in the new datasets. This is reflected in the numbers of opinion errors in en-Phone and zh-FoodBeverage, which are 125 (37.31%) and 395 (49.94%), respectively, exceeding the corresponding aspect errors of 99 (29.55%) and 246 (31.10%). Removing samples with implicit opinions reduces the opinion errors to 102 and 260 in en-Phone and zh-FoodBeverage, indicating that explicit opinion errors are slightly larger than explicit aspect errors. (4) The percentage of sentiment errors is relatively small, demonstrating the effectiveness of our proposed sentiment-specific horns tagging schema.

6 Related Work

ABSA Benchmark Datasets are mainly provided by the SemEval'14-16 shared challenges (Pontiki et al., 2014, 2015, 2016). The initial task is only to identify opinions expressed about specific entities and their aspects. In order to investigate more tasks, such as AOPE, E2E-ABSA, ASTE, TASD, and ASQP, researchers have re-annotated the datasets and constructed some new ones (Fan et al., 2019; Li



(a) Percentage of errors

Type	Example
Category	Input: the screen looked great. Gold: (DISPLAY#GENERAL, screen, great, POS) Pred.: (DISPLAY#DESIGN_FEATURES, screen, great, POS)
Aspect	Input: works flawlessly and decent battery life. Gold: (BATTERY#OPERATION_PERFORMANCE, battery, decent, POS) Pred.: (BATTERY#OPERATION_PERFORMANCE, battery life, decent, POS)
Opinion	Input: the keyboard is backlit and big enough for my fingers. Gold: (KEYBOARD#DESIGN_FEATURES, keyboard, NULL, POS) Pred.: (KEYBOARD#DESIGN_FEATURES, keyboard, big, POS)
Sentiment	Input: it starts up, runs without issues. Gold: (OS#OPERATION_PERFORMANCE, starts up, NULL, NEU) Pred.: (OS#OPERATION_PERFORMANCE, starts up, NULL, POS)

(b) Typical error examples from Laptop-ACOS.

Figure 2: Error analysis and case study. Though the predicted aspect and opinion differ from the golden ones in the above examples, they seem correct.

et al., 2019a; Xu et al., 2020; Wan et al., 2020; Cai et al., 2021). However, re-annotated datasets still contain the following limitations: (1) The data is collected from only one source, limiting the scope of the data; (2) the data size is usually small, where the maximum one is only around 4,000; (3) there is only a labeled quadruple per sentence and many samples share a common aspect, which makes the task easier; (4) the available public datasets are all in English. The shortcomings of existing benchmark datasets motivate us to crawl and curate more data from more domains, covering more languages and with higher quadruple density.

ASQP aims to predict the four sentiment elements to provide a complete aspect-level sentiment structure (Cai et al., 2021; Zhang et al., 2021a). The task is extended to several variants, e.g., capturing the quadruple of holder-target-expression-polarity (R et al., 2022; Lu et al., 2022) or the quadruple of target-aspect-opinion-sentiment in a dialogue (Li et al., 2022). Existing studies can be divided into the pipeline or generation paradigm. A typical *pipeline*-based work (Cai et al., 2021) has investigated different techniques to solve the subtasks accordingly. It consits of (1) first exploiting double propagation (DP) (Qiu et al., 2011) or JET (Xu et al., 2020) to extract the aspect-opinion-sentiment triplets and after that, detecting the aspect category to output the final quadruples; (2) first utilizing TAS-BERT (Wan et al., 2020) and the Extract-Classify scheme (Wang et al., 2017) to perform the aspect-opinion co-extraction and predicting category-sentiment afterward. Most studies fall in the *generation paradigm* (Zhang et al., 2021a; Mao et al., 2022; Bao et al., 2022; Gao et al., 2022). Zhang et al. (2021a) is the first generation-based method to predict the sentiment quads in an end-to-end manner via a *PARAPHRASE* modeling paradigm. It has been extended and overcome by Seq2Path (Mao et al., 2022) or tree-structure generation (Mao et al., 2022; Bao et al., 2022) to tackle the generation order issue or capture more information. Prompt-based generative methods are proposed to assemble multiple tasks as LEGO bricks to attain task transfer (Gao et al., 2022) or tackle few-shot learning (Varia et al., 2022). GEN-SCL-NAT (Peper and Wang, 2022) is introduced to exploit supervised contrastive learning and a new structured generation format to improve the naturalness of the output sequences for ASQP. However, existing methods either yield error propagation in the pipeline-based methods or slow computation in the generation-based methods. The shortcomings of existing methods motivate us to propose One-ASQP.

7 Conclusions

In this paper, we release two new datasets, with the first dataset being in Chinese, for ASQP and propose One-ASQP, a method for predicting sentiment quadruples simultaneously. One-ASQP utilizes a token-pair-based 2D matrix with sentiment-specific horns tagging, which allows for deeper interactions between sentiment elements, enabling efficient decoding of all aspect-opinion-sentiment triplets. An elaborately designed “[NULL]" token is used to identify implicit aspects or opinions effectively. Extensive experiments are conducted to demonstrate the effectiveness and efficiency of One-ASQP. Notably, existing strong baselines exhibit a decay in performance on the newly released datasets. We hope these datasets and One-ASQP will inspire further technical development in this area.

Acknowledgments

The work was partially supported by the IDEA Information and Super Computing Centre (ISCC) and the National Nature Science Foundation of China (No. 62201576).

Limitations

Our proposed One-ASQP still contains some limitations:

- Our One-ASQP does not solve the case of IA&IO. We defer the technical exploration of this issue to future work.
- One-ASQP has to split the ASQP task into two subtasks, ADC and AOSC. It is still promising to explore more effective solutions, e.g., by only one task, which can absorb deeper interactions between all elements.
- Generally, One-ASQP suffers more opinion errors than other sentiment elements due to the fine-grained annotation and implicit opinions issues. It is possible to tackle it by exploring more advanced techniques, e.g., syntax or semantics augmentation, to dig out deeper connections between options and other sentiment elements.
- One-ASQP tends to make errors when there are many aspect categories with small labeled quadruples. It is also significant to explore more robust solutions to detect the aspect categories in the low-resource scenario.
- Though we have released datasets in both English and Chinese, we do not explore ASQP in the multi-lingual scenario. We leave this as future work.

Ethics Statement

We follow the ACL Code of Ethics. In our work, there are no human subjects and informed consent is not applicable.

References

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

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 833–843. International Committee on Computational Linguistics.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 340–350. Association for Computational Linguistics.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12666–12674. AAAI Press.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 657–668. Association for Computational Linguistics.

Zehui Dai, Cheng Peng, Huajie Chen, and Yadong Ding. 2020. A multi-task incremental learning framework with category name embedding for aspect-category sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6955–6965. Association for Computational Linguistics.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2509–2518. Association for Computational Linguistics.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 7002–7012. International Committee on Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-

training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 388–397. Association for Computational Linguistics.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 504–515. Association for Computational Linguistics.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 397–406. Association for Computational Linguistics.

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, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1345–1359. Association for Computational Linguistics.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2022. Diaasq : A benchmark of conversational aspect-based sentiment quadruple analysis. *CoRR*, abs/2211.05705.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6714–6721. AAAI Press.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2886–2892. Association for Computational Linguistics.

Yangming Li, Lemao Liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019b. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4589–4599. Association for Computational Linguistics.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. 2021. Solving aspect category sentiment analysis as a text generation task. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4406–4416. Association for Computational Linguistics.

Xinyu Lu, Mengjie Ren, Yaojie Lu, and Hongyu Lin. 2022. ISCAS at semeval-2022 task 10: An extraction-validation pipeline for structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 1305–1312. Association for Computational Linguistics.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2215–2225. Association for Computational Linguistics.

Rajdeep Mukherjee, Tapas Nayak, Yash Butala, Sourangshu Bhattacharya, and Pawan Goyal. 2021. PASTE: A tagging-free decoding framework using pointer networks for aspect sentiment triplet extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9279–9291. Association for Computational Linguistics.

Joseph J. Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. *CoRR*, abs/2211.07743.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016.

Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 19–30. The Association for Computer Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 486–495. The Association for Computer Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

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

Raghav R, Adarsh Vemali, and Rajdeep Mukherjee. 2022. Etms@iitkgp at semeval-2022 task 10: Structured sentiment analysis using A generative approach. *CoRR*, abs/2205.00440.

Yuming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11285–11293. AAAI Press.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2022. Instruction tuning for few-shot aspect-based sentiment analysis. *CoRR*, abs/2210.06629.

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 *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9122–9129. AAAI Press.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3316–3322. AAAI Press.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1572–1582. International Committee on Computational Linguistics.

Meixi Wu, Wenya Wang, and Sinno Jialin Pan. 2020. Deep weighted maxsat for aspect-based opinion extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5618–5628. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2339–2349. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2416–2429. Association for Computational Linguistics.

Haiqin Yang, Xiaoyuan Yao, Yiqun Duan, Jianping Shen, Jie Zhong, and Kun Zhang. 2021. Progressive open-domain response generation with multiple controllable attributes. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3279–3285. ijcai.org.

Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. *IEEE ACM Trans. Audio Speech Lang. Process.*, 27(1):168–177.

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 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9209–9219. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based

sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 504–510. Association for Computational Linguistics.

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

A More Details about Datasets Construction

This section provides more details about constructing the two datasets, en-Phone and zh-FoodBeverage.

A.1 Data Sources

The English ASQP dataset, en-Phone, is collected from reviews on Amazon UK [3], Amazon India [4] and Shopee [5] in July and August of 2021, covering 12 cell phone brands, such as Samsung, Apple, Huawei, OPPO, Xiaomi, etc.

The first Chinese ASQP dataset, zh-FoodBeverage, is collected from the Chinese comments on forums [6], Weibo [7], news [8] and e-commerce platforms [9] in the years 2019-2021 under the categories of Food and Beverage.

A.2 Annotation Guidelines

The following outlines the guidelines for annotating the four fundamental sentiment elements of ASQP and their outcomes. It can be noted that our labeled ASQP quadruples are more fine-grained and more difficult than those in existing ASQP benchmark datasets.

A.2.1 Aspect Categories

The **aspect category** defines the type of the concerned aspect. Here, we apply a two-level category system, which is defined by our business experts for the sake of commercial value and more detailed information. For example, "Screen" is a first-level category. It can include second-level categories, such as "Clarity", "General", and "Size", to form the final second-level categories as "Screen#Clarity", "Screen#General", and "Screen#Size". In the experiments, we only consider the second-level categories.

As reported in Table 2, the number of categories for en-Phone and zh-FoodBeverage is 88 and 138, respectively. The number of labeled quadruples per category is larger than 5. Though Laptop-ACOS consists of 121 categories, if we filter out the categories with less than 5 annotated quadruples, the number of categories is reduced to 75. Hence, we provide more dense and rich datasets for ASQP.

A.2.2 Aspect Terms

The **aspect** term is usually a noun or a noun phrase, indicating the opinion target, in the text. It can be implicit in a quadruple (Cai et al., 2021). For the sake of commercial analysis, we exclude sentences without aspects. Moreover, to provide more fine-grained information, we include three additional rules:

- The aspect term can be an adjective or verb when it can reveal the sentiment categories. For example, as the example of en-Phone in Table 10, "recommended" is also labeled as an aspect in "Highly recommended" because it can identify the category of "Buyer_Atitude#Willingness_Recommend". In Ex. 1 and Ex. 4 of Table 9, "clear" and "cheap" are labeled as the corresponding aspect terms because they can specify the category of "Screen#Clarity" and "Price#General", accordingly.
- Pronoun is not allowed to be an aspect term as it cannot be identified by the quadruples only. For example, in the example of "prettyyyy and affordable too!!! I love it!! Thankyouuu!!", "it" cannot be labeled as the aspect though we know it indicates a phone from the context.
- Top-priority in labeling fine-grained aspects. For example, in the example of "Don't purchase this product", "purchase" is more related to a customer's purchasing willingness while "product" is more related to the overall comment, we will label "purchase" as the aspect.

A.2.3 Opinion Terms

The **opinion** term describes the sentiment towards the aspect. An opinion term is usually an adjective or a phrase with sentiment polarity. Here, we

[3] https://www.amazon.co.uk/
[4] https://www.amazon.in/
[5] https://shopee.com.my/
[6] http://foodmate.net/
[7] https://weibo.com/
[8] https://chihe.sohu.com/
[9] https://www.jd.com/,https://www.taobao.com/

	Sentence	Labeled Quadruples
Ex. 1	This screen is good overall, although the screen size is not large, but looks very clear	(Screen#General, screen, good overall, POS) (Screen#Size, screen size, not large, NEG) (Screen#Clarity, clear, very, POS)
Ex. 2	Don't like face recognition and battery life.	(Security#Screen Unlock, face recognition, Don't like, NEG) (Battery/Longevity#Battery life, battery life, Don't like, NEG)
Ex. 3	Very fast delivery & phone is working well.	(Logistics#Speed, delivery, Very fast, POS) (Overall Rating#General, phone, working well, POS)
Ex. 4	It's very cheap. The first time I bought the phone I wanted.	(Price#General, cheap, very, POS)

Table 9: Three opinionated sentences and the labeled quadruples.

Source	Opinionated sentences	Quadruples
en-Phone	Item received took a long time. Looks nice and good quality too. Price is cheaper than retail costs. Bought it for my mom and she likes it! Highly recommended.	(Logistics#Logistics speed, Item received, took a long time, NEG), (Exterior Design#Aesthetics, Looks, nice, POS), (Product Quality#General, quality, good, POS), (Price#General, Price, cheaper than retail costs, POS), (Audience#Users, mom, likes, POS), (Buyer attitude#Recommendable, recommended, Highly, POS)
zh-FoodBeverage	我挑选这个品牌的主要原因是它这里含一项乳铁蛋白促进宝宝吸收的，所以宝宝喝它没有奶瓣的情况，同时它的口味也接近母乳，宝宝很爱喝。可惜的是，买了那么多奶粉，赠品没收到，也不知道哪个环节出的问题，连赠品的影子都没看见，只能认倒霉了，无处对证去 The main reason I picked this brand is that it contains Lactoferrin to promote baby's absorption, so the baby drinks it without a milk valve, while its taste is also close to breast milk, the baby loves to drink. Unfortunately, I bought so much milk powder and did not receive a gift, and I do not know which part of the problem, even the shadow of the gift did not see, can only admit bad luck, there is no evidence to testify.	(成分#营养成分, 乳铁蛋白, 含, POS), (营养#吸收, 宝宝吸收, 促进, POS), (不良反应#其他不适, 奶瓣, 没有, POS), (味道#综合味道, 口味, 接近母乳, POS), (使用#受众群体, 宝宝, 很爱喝, POS), (促销#赠品, 赠品, 没收到, NEG), (促销#赠品, 赠品, 没看见, NEG), (Ingredients#Nutritional Composition, lactoferrin, contains, POS), (Nutrition#Absorption, baby's absorption, promote, POS), (Adverse reactions#Other discomfort, milk valve, without, POS), (Flavor#Comprehensive flavor, taste, close to breast milk, POS), (Use#Audience group, baby, loves to drink, POS), (Promotion#Giveaway, gift, did not receive, NEG), (Promotion#Giveaway, gift, did not see, NEG),

Table 10: Typical examples of the labeled quadruples in en-Phone and zh-FoodBeverage.

include more labeling criteria:

- When there is a negative word, e.g., "Don't", "NO", "cannot", "doesn't", the negative word should be included in the opinion term. For example, "not large" and "Don't like" are labeled as the corresponding opinion terms in Ex. 1 and Ex. 2 of Table 9
- When there is an adverb or a preposition, e.g., "very", "too", "so", "inside", "under", "outside", the corresponding adverb or preposition should be included in the opinion term. For example, in Ex. 3 of Table 9, "Very fast" is labeled as an opinion term. Usually, in Restaurant-ACOS and Laptop-ACOS, "Very" is not included in the opinion term. Moreover, in Ex. 1 of Table 9, "very" in "very clear" is labeled as an opinion term while in Ex. 4, "very" in "very cheap" is labeled as the opinion term.

These examples show that our labeled opinion terms are more fine-grained and complicated, but more valuable for real-world applications. This increases the difficulty of extracting opinion terms and demonstrates the significance of our released datasets to the ASBA community.

A.2.4 Sentiment Polarity

The **sentiment polarity** belongs to one of the sentiment classes, {POS, NEU, NEG}, for the positive, neutral, and negative sentiment, respectively. In zh-FoodBeverage, for commercial considerations, we only label sentences with positive and negative sentiments and exclude those with neutral sentiment.

A.3 Quadruple Density Analysis

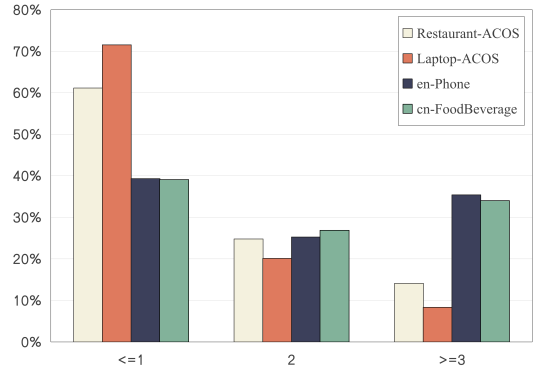


Figure 3: The ratio of the number of quadruples per sentence in four datasets.

For better illustration, we count the number

Method	Restaurant-ACOS		Laptop-ACOS		en-Phone		zh-FoodBeverage	
	Train	Inference	Train	Inference	Train	Inference	Train	Inference
Extract-Classify	38.43	14.79	72.25	20.23	158.34	25.23	301.42	70.34
Paraphrase	30.52	58.23	59.23	69.23	99.23	160.56	664.23	673.32
GEN-SCL-NAT	35.32	61.64	63.53	72.23	104.23	175.23	748.56	706.43
OneASQP (base)	11.23	6.34 (29.35)	19.03	8.34 (39.83)	32.23	6.32 (35.45)	71.23	13.23 (31.74)
OneASQP (large)	17.63	14.63 (44.62)	36.63	8.45 (49.45)	105.23	10.34 (61.23)	140.23	30.46 (56.32)

Table 11: Time cost in seconds on all datasets. For a fair comparison with baselines, we record our One-ASQP inference time when setting the batch size to 1 and report them in the round bracket. The default batch size is 32.

of quadruples per sentence in four datasets and show the ratios in Fig. 3. It is shown that (1) In terms of sentences with at most one labeled quadruple, Restaurant-ACOS contains 61.12% of the sentences and it is 71.54% in Laptop-ACOS. Meanwhile, it is 39.33% and 39.10% in en-Phone and zh-FoodBeverage, respectively. (2) In terms of sentences with at least three labeled quadruples, it drops significantly to 14.09% in Restaurant-ACOS and 8.34% in Laptop-ACOS. Meanwhile, it is 35.19% in en-Phone and 34.01% in zh-FoodBeverage. Hence, our released datasets are more dense and balanced.

B More Experimental Results

B.1 Computation Efficiency

Table 11 reports the time cost (in seconds) on all four datasets. The base versions of the corresponding LMs are applied in Extract-Classify. It shows that One-ASQP is efficient in both training and inference, which is a favorite for real-world deployment.

	Variant 1	Variant 2	One-ASQP
Restaurant-ACOS	58.39	57.23	**59.78**
Laptop-ACOS	41.05	39.12	**41.37**
en-Phone	51.23	49.72	**53.58**
zh-FoodBeverage	57.23	55.95	**57.79**

Table 12: Comparison of One-ASQP with two other variants for ASQP.

B.2 Effect of Variants of Interactions

Though our One-ASQP separates the task into ACD and AOSC. There are still other variants to resolve the ASQP task. Here, we consider two variants:

Variant 1: The ASQP task is separated into three sub-tasks: aspect category detection (ACD), aspect-opinion pair extraction (AOPC), and sentiment detection. More specifically, ACD and sentiment detection are fulfilled by classification models. For AOPC, we adopt the sentiment-specific horns tagging schema proposed in Sec. 3.2.2. That is, we only co-extract the aspect-opinion pairs. In the implementation, we set the tags of AB-OE-*SENTIMENT to AB-EO and reduce the number of tags for AOSC to three, i.e., {AB-OB, AE-OE, AB-OE}.

Variant 2: We solve the ASQP task by a unified framework. Similarly, via the sentiment-specific horns tagging schema proposed in Sec. 3.2.2, we extend the tags of AB-EO-*SENTIMENT to AB-OE-*SENTIMENT-*CATEGORY. Hence, the number of tags increases from 5 to $2 + |\mathcal{S}| * |\mathcal{C}|$, where $|\mathcal{S}|$ is the number of sentiment polarities and $|\mathcal{C}|$ is the number of categories. This setting allows us to extract the aspect-opinion pairs via the 2D matrix while decoding the categories and sentiment polarities via the tags.

Tabel 12 reports the compared results on four datasets, where the base versions of the corresponding LMs are applied. The results show that (1) our One-ASQP performs the best over the proposed two variants. We conjecture that the aspect-opinion-sentiment triplets are in a suitable tag space and our One-ASQP can absorb their interactions effectively. (2) Variant 2 performs the worst among all results. We conjecture that the search tag space is too large and the available datasets do not contain enough information to train the models.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
0, 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2, A

☑ B1. Did you cite the creators of artifacts you used?
1, 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
0

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
2, A

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2, A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2, 4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
2, 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

2, A

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
A

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
A

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
2

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.