Aspect Sentiment Quad Prediction as Paraphrase Generation *

Wenxuan Zhang[1], Yang Deng[1], Xin Li[2], Yifei Yuan[1], Lidong Bing[2] and Wai Lam[1]

[1]The Chinese University of Hong Kong

[2]DAMO Academy, Alibaba Group

{wxzhang,ydeng,yfyuan,wlam}@se.cuhk.edu.hk

{xinting.lx,l.bing}@alibaba-inc.com

Abstract

Aspect-based sentiment analysis (ABSA) has been extensively studied in recent years, which typically involves four fundamental sentiment elements, including the aspect category, aspect term, opinion term, and sentiment polarity. Existing studies usually consider the detection of partial sentiment elements, instead of predicting the four elements in one shot. In this work, we introduce the Aspect Sentiment Quad Prediction (ASQP) task, aiming to jointly detect all sentiment elements in quads for a given opinionated sentence, which can reveal a more comprehensive and complete aspect-level sentiment structure. We further propose a novel PARAPHRASE modeling paradigm to cast the ASQP task to a paraphrase generation process. On one hand, the generation formulation allows solving ASQP in an end-to-end manner, alleviating the potential error propagation in the pipeline solution. On the other hand, the semantics of the sentiment elements can be fully exploited by learning to generate them in the natural language form. Extensive experiments on benchmark datasets show the superiority of our proposed method and the capacity of cross-task transfer with the proposed unified PARAPHRASE modeling framework.

1 Introduction

As a fine-grained opinion mining problem, aspect-based sentiment analysis (ABSA) aims to analyse sentiment information at the aspect level (Liu, 2012; Pontiki et al., 2014). Typically, four fundamental sentiment elements are involved in ABSA, including 1) *aspect category* denoting the type of the concerned aspect; 2) *aspect term* which can be either explicitly or implicitly mentioned in the given text; 3) *opinion term* which describes the opinion towards the aspect; and 4) *sentiment polarity* denoting the sentiment class. Given an example sentence "*The pasta is over-cooked!*", the sentiment elements are "*food quality*", "*pasta*", "*over-cooked*", and "*negative*", respectively.

Due to its broad application scenarios, many research efforts have been made on ABSA to predict or extract those sentiment elements (Pontiki et al., 2014, 2015, 2016). Early studies focus on the prediction of a single element such as aspect term extraction (Liu et al., 2015; Xu et al., 2018), aspect category detection (Zhou et al., 2015), aspect sentiment classification based on either an aspect category (Ruder et al., 2016; Hu et al., 2019a) or an aspect term (Huang and Carley, 2018). More recent works propose to extract multiple associated sentiment elements at the same time (Zhang et al., 2021). For example, Chen et al. (2020) consider the aspect and opinion term pairwise extraction; Peng et al. (2020) propose the aspect sentiment triplet extraction (ASTE) task to detect the (aspect term, opinion term, sentiment polarity) triplets; Wan et al. (2020) handle the target aspect sentiment detection (TASD) task that jointly detects the aspect category, aspect term, and sentiment polarity.

Despite their popularity, these ABSA tasks only attempt to perform partial prediction instead of providing a complete aspect-level sentiment picture, *i.e.*, identifying the four sentiment elements in one shot. To this end, we introduce the aspect sentiment quad prediction (**ASQP**) task, aiming to predict all (aspect category, aspect term, opinion term, sentiment polarity) quads for a given opinionated sentence. This new task compensates for the drawbacks of previous tasks and helps us comprehensively understand user's aspect-level opinions.

To tackle ASQP, one straightforward idea is to decouple the quad prediction problem into several sub-tasks and solve them in a pipeline manner. However, such multi-stage approaches would suffer severely from error propagation because the overall

prediction performance hinges on the accuracy of every step (Peng et al., 2020; Chen et al., 2020). Besides, the involved sub-tasks, which are usually formulated as either token-level or sequence-level classification problems, underutilize the rich semantic information of the label (*i.e.*, the meaning of sentiment elements to be predicted) since they treat the labels as number indices during training. Intuitively, the aspect term "*pasta*" is unlikely to be coupled with the aspect category "*service general*" due to the large semantic gap between them. But such information cannot be suitably utilized in those classification-type methods.

Inspired by recent success in formulating various NLP tasks as text generation problems (Athiwaratkun et al., 2020; Paolini et al., 2021; Liu et al., 2021), we propose to tackle ASQP in a sequence-to-sequence (S2S) manner in this paper. On one hand, the sentiment quads can be predicted in an end-to-end manner, alleviating the potential error propagation in the pipeline solutions. On the other hand, the rich label semantic information could be fully exploited by learning to generate the sentiment elements in the natural language form.

Exploiting generation modeling for the ASQP task mainly faces two challenges: (i) how to linearize the desired sentiment information so as to facilitate the S2S learning? (ii) how can we utilize the pretrained models for tackling the task, which is a common practice now for solving various ABSA tasks (Xu et al., 2020; Cai et al., 2020)? To handle these two challenges, we propose a novel PARAPHRASE modeling paradigm, which transforms the ASQP task as a paraphrase generation problem (Bhagat and Hovy, 2013). Specifically, our approach linearizes the sentiment quad into a natural language sentence as if we were paraphrasing the input sentence and highlighting its major sentiment elements. For example, we can transform the aforementioned sentiment quad (*food quality*, *pasta*, *over-cooked*, *negative*) to a sentence "*Food quality is bad because pasta is over-cooked*". Such a linearized target sequence, paired with the input sentence "*The pasta is over-cooked!*" can then be used to learn the mapping function of a generation model. We can seamlessly utilize the large pretrained generative models such as T5 (Raffel et al., 2020) by fine-tuning with such input-target pairs. Therefore, the rich label semantics of the sentiment elements is naturally fused with the rich knowledge of the pretrained models in the form of natural sentences, rather than directly treating the desired sentiment quad text sequence as the generation target (Zhang et al., 2021).

We summarize our contributions as follows: 1) We study a new task, namely aspect sentiment quad prediction (ASQP) in this work and introduce two datasets with sentiment quad annotations for each sample, aiming to analyze more comprehensive aspect-level sentiment information. 2) We propose to tackle ASQP as a paraphrase generation problem, which can predict the sentiment quads in one shot and fully utilize the semantics information of natural language labels. 3) Extensive experiments show that the proposed PARAPHRASE modeling is effective to tackle ASQP as well as other ABSA tasks, outperforming the previous state-of-the-art models in all cases. 4) The experiment also suggests that our PARAPHRASE method naturally facilitates the knowledge transfer across related tasks with the unified framework, which can be especially beneficial in the low-resource setting.[1]

2 Related Work

ABSA has been extensively studied in recent years where the main research line is the extraction of the sentiment elements. Early studies focus on the prediction of a single element such as extracting the aspect term (Liu et al., 2015; Yin et al., 2016; Xu et al., 2018; Ma et al., 2019), detecting the mentioned aspect category (Zhou et al., 2015; Bu et al., 2021), and predicting the sentiment polarity, given either an aspect term (Wang et al., 2016; Huang and Carley, 2018; Zhang and Qian, 2020) or an aspect category (Ruder et al., 2016; Hu et al., 2019a). Some works further consider the joint detection of two sentiment elements, including the pairwise extraction of aspect and opinion term (Wang et al., 2017; Chen et al., 2020; Zhao et al., 2020); the prediction of aspect term and its corresponding sentiment polarity (Li et al., 2019a; He et al., 2019; Hu et al., 2019b; Luo et al., 2019; Chen and Qian, 2020); and the co-extraction of aspect category and sentiment polarity (Cai et al., 2020).

More recently, triplet prediction tasks are proposed in ABSA, aiming to predict the sentiment elements in triplet format. Peng et al. (2020) propose the aspect sentiment triplet extraction (ASTE) task, which has received lots of attention (Xu et al., 2020; Huang et al., 2021; Mao et al., 2021; Chen

---

et al., 2021). Wan et al. (2020) introduce the target aspect sentiment detection (TASD) task, aiming to predict the aspect category, aspect term, and sentiment polarity simultaneously, which can handle the case where the aspect term is implicit expressed in the given text (treated as "null") (Wu et al., 2021). Built on top of those tasks, we introduce the aspect sentiment quad prediction problem, aiming to predict the four sentiment elements in one shot, which can provide a more detailed and comprehensive sentiment structure for a given text.

Adopting pretrained transformer-based models such as BERT (Devlin et al., 2019) has become a common practice for tackling the ABSA problem. Especially, many ABSA tasks benefit from appropriately utilizing the pretrained models. Sun et al. (2019) transform the aspect sentiment classification task as a language inference problem by constructing an auxiliary sentence. Chen et al. (2021) and Mao et al. (2021) formulate multiple ABSA tasks as a reading comprehension task to fully utilize the knowledge of the pre-trained model. Very recently, there are some attempts on tackling ABSA problem in a S2S manner, either treating the class index (Yan et al., 2021) or the desired sentiment element sequence (Zhang et al., 2021) as the target of the generation model. In this work, we propose a PARAPHRASE modeling that can better utilize the knowledge of the pre-trained model via casting the original task to a paraphrase generation process.

3 Methodology

3.1 Problem Statement

Given a sentence x, aspect sentiment quad prediction (ASQP) aims to predict all aspect-level sentiment quadruplets $\{(c, a, o, p)\}$ which corresponds to the aspect category, aspect term, opinion term, and sentiment polarity, respectively. The aspect category c falls into a category set V_c; the aspect term a and the opinion term o are typically text spans in the sentence x while the aspect term can also be null if the target is not explicitly mentioned: $a \in V_x \cup \{\varnothing\}$ and $o \in V_x$ where V_x denotes the set containing all possible continuous spans of x. The sentiment polarity p belongs to one of the sentiment class $\{$POS, NEU, NEG$\}$ denoting the positive, neutral, and negative sentiment respectively.

3.2 ASQP as Paraphrase Generation

We propose a PARAPHRASE modeling paradigm to transform the ASQP task as a paraphrase genera-
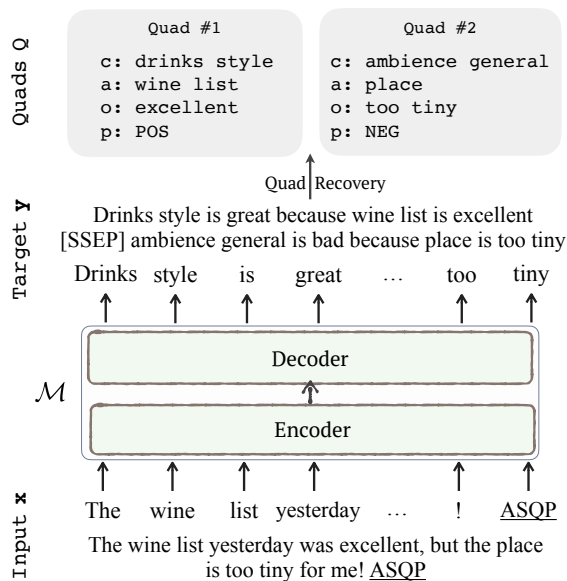

Figure 1: Overview of the paraphrase generation framework. The underlined task identifier in the input is only used under the cross-task transfer setting.

tion problem and solve it in a sequence-to-sequence manner. As depicted in Figure 1, given a sentence x, we aim to generate a target sequence y with an encoder-decoder model $\mathcal{M} : x \rightarrow y$ where y contains all the desired sentiment elements. Then the sentiment quads $Q = \{(c, a, o, p)\}$ can be recovered from y for making the prediction.

On one hand, the semantics of the sentiment elements in Q could be fully exploited by generating them in the natural language form in y. On the other hand, the input and target are both natural language sentences, which can naturally utilize the rich knowledge in the pretrained generative model.

PARAPHRASE Modeling To facilitate the S2S learning, given the sentence label pair (x, Q), an important component of the PARAPHRASE modeling framework is to linearize the sentiment quads Q to a natural language sequence y for constructing the input target pair (x, y).

Ideally, we aim to neglect unnecessary details in the input sentence while highlight the major sentiment elements in the target sentence during the paraphrasing process. Based on this motivation, we linearize a sentiment quad $q = (c, a, o, p)$ to a natural sentence as follows:

$$\mathcal{P}_c(c) \text{ is } \mathcal{P}_p(p) \text{ because } \mathcal{P}_a(a) \text{ is } \mathcal{P}_o(o).$$

where $\mathcal{P}_z(\cdot)$ is the projection function for $z \in \{c, a, o, p\}$, which maps the sentiment element z from the original format to a natural language form.

By adopting suitable projection functions, a structured sentiment quad q can then be transformed to an equivalent natural language sentence.

For the input sentence $\boldsymbol{x}$ with multiple sentiment quads, we first linearize each quad q to a natural sentence as described above. Then these sentences are concatenated with a special symbol `[SSEP]` to form the final target sequence $\boldsymbol{y}$, containing all the sentiment quads for the given sentence.

Target Construction for ASQP Since the aspect category c and opinion term o in each sentiment quad are already in the natural language form, their projection functions just keep the original formats: $\mathcal{P}_c(c) = c$ and $\mathcal{P}_o(o) = o$. For the sentiment polarity, the projection is as follows:

$$\mathcal{P}_p(p) = \begin{cases} great & \text{if } p = \text{POS} \\ ok & \text{if } p = \text{NEU} \\ bad & \text{if } p = \text{NEG} \end{cases} \quad (1)$$

where the main idea is to transform the sentiment label from the original class format to a natural language expression and also ensure the coherence of the whole linearized target sequence so that the semantics of the sentiment polarity can be exploited by the generation model. Note that the specific mapping can either be pre-defined with commonsense knowledge as in Equation 1 or dataset-dependent which utilizes the most common concurring opinion term for each sentiment polarity as the sentiment expression.

As for the aspect term, we map it to an implicit pronoun if it is not explicitly mentioned, otherwise we can just use the original natural language form:

$$\mathcal{P}_a(a) = \begin{cases} it & a = \varnothing \\ a & otherwise \end{cases} \quad (2)$$

This is to mimic the writing process where users often use a pronoun such as "*it*" or "*this*" to refer to a target that is not explicitly expressed.

After defining the specific projection functions for each sentiment element, we can then transform a sentiment quad to a sentence containing all the elements in the natural language form to facilitate the S2S learning. Two target construction examples for the ASQP task are shown in Figure 2.

3.3 Sequence-to-Sequence Learning

The input-to-target generation can be modeled with a classical encoder-decoder model such as the

Input-1	*The pasta yesterday was delicious!*
Label-1 ⇓	(c, a, o, p): (food quality, pasta, delicious, POS) ⇓
Target-1	Food quality is great because pasta is delicious
Input-2	*Everything they serve here ... was just very disappointed, I wish they would change next time*
Label-2 ⇓	(c, a, o, p): (food quality, NULL, disappointed, NEG) ⇓
Target-2	Food quality is bad because it is disappointed

Figure 2: Two examples of the target sentence construction for the ASQP task. Better viewed in colors.

Transformer architecture (Vaswani et al., 2017). Given the sentence $\boldsymbol{x}$, the encoder first transforms it into a contextualized encoded sequence $\boldsymbol{e}$. The decoder then aims to model the conditional probability distribution of the target sentence $\boldsymbol{y}$ given the encoded input representation: $p_\theta(\boldsymbol{y}|\boldsymbol{e})$ which is parameterized by θ.

At the i-th time step, the decoder output $\boldsymbol{y}_i$ is computed based on both the encoded input $\boldsymbol{e}$ and the previous outputs $\boldsymbol{y}_{<i}$: $\boldsymbol{y}_i = f_{dec}(\boldsymbol{e}, \boldsymbol{y}_{<i})$ where $f_{dec}(\cdot)$ denotes the decoder computations. To obtain the probability distribution for the next token, a softmax function is then applied:

$$p_\theta(\boldsymbol{y}_{i+1}|\boldsymbol{e}, \boldsymbol{y}_{<i+1}) = \text{softmax}(W^T \boldsymbol{y}_i) \quad (3)$$

where W maps the prediction $\boldsymbol{y}_i$ to a logit vector, which can then be used to compute the probability distribution over the whole vocabulary set.

Training With a pretrained encoder-decoder model such as T5 (Raffel et al., 2020), we can initialize θ with the pretrained parameter weights and further fine-tune the parameters on the input-target pair to maximize the log-likelihood $p_\theta(\boldsymbol{y}|\boldsymbol{e})$:

$$\max_\theta \log p_\theta(\boldsymbol{y}|\boldsymbol{e}) = \sum_{i=1}^{n} \log p_\theta(\boldsymbol{y}_i|\boldsymbol{e}, \boldsymbol{y}_{<i}) \quad (4)$$

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

Inference and Quad Recovery After the training, we generate the target sequence $\boldsymbol{y}'$ in an autoregressive manner and select the token with the highest probability over the vocabulary set as the next token at each time step. Then we can recover the predicted sentiment quads Q' from the generations. Specifically, we first split the possible multiple quads via detecting the pre-defined separation token `[SSEP]`. Then for each linearized sentiment quad sequence, we extract the sentiment elements according to the modeling strategy introduced in

Sec 3.2 and compare them with the gold sentiment quad in Q for the evaluation. If such decoding fails, for example, the generated sequence violates the defined format, we treat the prediction as null.

3.4 ABSA as Paraphrase Generation

The proposed PARAPHRASE modeling in fact provides a general paradigm to tackle the ABSA problem, which transforms the sentiment element prediction to a paraphrase generation process. Therefore, it can be easily extended to handle other ABSA tasks as well: we only need to change the projection functions for each sentiment element to suit the need for each task. We take the target aspect sentiment detection (TASD) (Wan et al., 2020) and aspect sentiment triplet extraction (ASTE) (Peng et al., 2020) tasks as two examples here[2].

The TASD task predicts the (c, a, p) triplets where all sentiment elements have the same condition as in the ASQP problem. Since it does not involve the opinion term prediction, we just let $\mathcal{P}_o(o) = \mathcal{P}_p(p)$ which uses a manually constructed opinion word as the opinion expression to describe the sentiment in the paraphrase. Other projection functions can remain the same as in the ASQP task. For instance, it transforms the (*service general*, *waiter*, NEG) triplet to the target sentence "*Service general is bad because waiter is bad*".

For the ASTE task aiming to predict (a, o, p) triplets, we map the aspect category to an implicit pronoun such as "*it*" ($\mathcal{P}_o(o) = it$) in all cases. Besides, it ignores the implicit aspect term, which means $a \in V_{\boldsymbol{x}}$. We then always use the aspect term in its original natural language form: $\mathcal{P}_a(a) = a$. Given an example triplet (*Chinese food*, *nice*, POS), a target sentence "*It is great because Chinese food is nice*" can be constructed accordingly.

3.5 Cross-task Knowledge Transfer

In practice, it is usually rather difficult and expensive to collect large-scale annotated data for complex ABSA problems like ASQP. Fortunately, as introduced in the last section, the proposed PARAPHRASE method tackles various ABSA tasks in a unified framework. This characteristic naturally enables the knowledge to be easily transferred across related ABSA tasks, which is especially beneficial

---

[2] In fact, any ABSA task involving the prediction of one or multiple sentiment elements can be considered as a sub-task of ASQP. We mainly discuss ASTE and TASD tasks in this paper since they are more closely related.

| | Rest15 | | | | Rest16 | | | |
	#S	#+	#0	#-	#S	#+	#0	#-
Train	834	1005	34	315	1264	1369	62	558
Dev	209	252	14	81	316	341	23	143
Test	537	453	37	305	544	583	40	176

Table 1: Data statistics for the ASQP task. #S, #+, #0, and #- denote the number of sentences, number of positive, neutral, negative quads respectively.

under the low-resource setting (*i.e.*, the labeled data for the concerned task is insufficient).

We investigate cross-task transfer for the concerned ASQP task, with the help of its two subtasks, including ASTE and TASD. Similar to recent works on using "prompt" as the task identifier (Raffel et al., 2020; Liu et al., 2021), we add a task-specific text suffix (*e.g.*, ASQP for the ASQP task in Figure 1) to the input sentence before feeding it to the model for specifying which task the model should perform. Since the PARAPHRASE paradigm provides a consistent training objective, the rich task-specific knowledge can first be learned from training on the TASD and ASTE tasks, and then naturally transferred to the ASQP task via fine-tuning on the (limited) ASQP data.

4 Experimental Setup

Dataset We build the ASQP datasets based on SemEval Shared Challenges (Pontiki et al., 2015, 2016). The annotations of the opinion term and aspect category are derived from Peng et al. (2020) and Wan et al. (2020) respectively. We align the samples from these two sources and merge the annotations with the same aspect term in each sentence as the anchor. We further conduct some additional annotations:

- Sentences without explicit aspect terms are ignored in Peng et al. (2020), we add these sentences back to our ASQP datasets and manually annotate the opinion terms for them, based on the given aspect category. For example, given a sentence "*Everything we had was good...*" with implicit aspect term, we then annotate "*good*" as the opinion term according to the aspect category "*food quality*". The quads with implicit opinion expressions are discarded.
- For the same aspect term associated with multiple aspect categories and/or opinion terms, the merging result will have more than four sentiment elements for each quad, we then manually check those cases to correct the labels to ensure

Type	Methods	Rest15			Rest16		
		Pre	Rec	F1	Pre	Rec	F1
Pipeline	HGCN-BERT + BERT-Linear	24.43	20.25	22.15	25.36	24.03	24.68
	HGCN-BERT + BERT-TFM	25.55	22.01	23.65	27.40	26.41	26.90
	TASO-BERT-Linear	41.86	26.50	32.46	49.73	40.70	44.77
Unified	TASO-BERT-CRF	44.24	28.66	34.78	48.65	39.68	43.71
	GAS	<u>45.31</u>	<u>46.70</u>	<u>45.98</u>	<u>54.54</u>	<u>57.62</u>	<u>56.04</u>
Ours	PARAPHRASE	**46.16**	**47.72**	**46.93**	**56.63**	**59.30**	**57.93**
	w/o sentiment polarity semantics	45.30	46.87	46.07	56.56	58.82	57.67
	w/o aspect category semantics	44.65	46.59	45.60	56.27	58.38	57.31
	w/o polarity & category semantics	43.46	45.19	44.30	56.04	57.53	56.77

Table 2: Main results of the ASQP task and ablations on label semantics for the proposed method. The best and second best results are in bold and underlined respectively. Scores are averaged over 5 runs with different seeds.

the aspect category and opinion term are matched in the same quad.

Every sample is annotated by two human annotators and the conflict cases would be checked. Finally, we obtain two datasets, namely Rest15 and Rest16, where each data instance contains a review sentence with one or multiple sentiment quads. We further split 20% of the data from the training set as the validation set. The statistics is summarized in Table 1.

Evaluation Metrics   We employ F1 scores as the main evaluation metrics. A sentiment quad prediction is counted as correct if and only if all the predicted elements are exactly the same as the gold labels. We also report the precision (Pre) and recall (Rec) scores for the ASQP task.

Experiment Details   The averaged scores over five runs with different random seed initialization are reported. We adopt the T5-BASE (Raffel et al., 2020) as the pretrained generative model described in Sec 3.3, which adopts a classical Transformer encoder-decoder network architecture. Regarding the training, we use a batch size of 16 and learning rate being 3e-4. The number of training epochs is 20 for all experiments. During the inference, we utilize greedy decoding for generating the output sequence. We also experiment with beam search decoding with the number of beams being 3, 5, and 8 respectively, all leading to similar performance with the greedy decoding. Therefore, greedy decoding is used for simplicity.

Baselines   Since the ASQP task has not been explored previously, we construct two types of baselines to compare with our PARAPHRASE method:

- **Pipeline model**: we cascade models in a pipeline manner for the quad prediction: **HGCN** (Cai et al., 2020) for jointly detecting the aspect category and sentiment polarity, followed by a BERT-based model extracting the aspect and opinion term (Li et al., 2019b), given the predicted aspect category and sentiment. The latter one can be either equipped with a linear layer (**BERT-Linear**) or a transformer block (**BERT-TFM**) on top.
- **Unified model**: we first modify TAS (Wan et al., 2020), a state-of-the-art unified model to extract (c, a, p) triplet, for tackling the ASQP task. TAS expands each original data sample into multiple samples, each with a specific aspect category and sentiment polarity pair, to solve the task in an end-to-end manner. We change its tagging schema to predict aspect and opinion term simultaneously for constructing a unified model to predict the quad, denoted as TASO (TAS with Opinion). There are two variants in terms of the prediction layer: either using a linear classification layer (**TASO-Linear**) or the CRF layer (**TASO-CRF**). We also consider a generation-type baseline **GAS**, originally proposed in (Zhang et al., 2021), we modify it to directly treat the sentiment quads sequence as the target for learning the generation model. It uses the same pretrained model as ours.

5   Results and Discussions

5.1   Main Results

The result for the ASQP task is reported in Table 2. There are some notable observations: Firstly, the performance of the pipeline methods is far from satisfactory. Although both adopting BERT as the backbone, the unified methods (*e.g.*, TASO-BERT-

Linear) perform much better than the pipeline ones (*e.g.*, HGCN-BERT + BERT-Linear). This verifies our assumption that the pipeline solutions tend to accumulate errors from the sub-task models and finally affect the performance of the final quad prediction. Secondly, among the unified methods, GAS outperforms two variants of TASO by a large margin, showing the effectiveness of the sequence-to-sequence modeling for the ASQP task. Besides, to solve the task in a unified manner, TASO expands the dataset to $|V_c| \times |V_p|$ times the original size, leading to large computation costs and training time. Thirdly, we can see that our proposed method, PARAPHRASE modeling achieves the best performance on all metrics across two datasets. Our method tackles the ASQP problem in an end-to-end manner, alleviating the possible error propagation in the pipeline solutions. Moreover, compared with the GAS method using the same pre-trained model, our PARAPHRASE also achieves superior results, suggesting that constructing target sequence in the natural language form is a better way for utilizing the knowledge from the pre-trained generative model, thus leading to better performance.

5.2 Effect of Label Semantics

Different from previous classification-type methods for tackling ABSA problem, our PARAPHRASE modeling can take advantage of the semantics of the sentiment elements by generating the natural language labels. We conduct ablation studies to further investigate the impact of the label semantics. Specifically, instead of mapping the label to the natural language form with the projection functions introduced in Sec 3.2, we map each label to a special symbol, similar as the number index in the classification-type models, for representing each label class. We consider three cases: (1) w/o sentiment polarity semantics: $\mathcal{P}_p(p_i) = \text{SP}i$ where p_i is a sentiment polarity, i denotes the index. For example, we map the positive class as $\text{SP}1$; (2) w/o aspect category semantics: $\mathcal{P}_c(c_j) = \text{AC}j$ where we project the aspect category c_j to a symbol with its index j[3]. For instance, the aspect category "food quality" will be mapped to $\text{AC}3$; (3) w/o polarity & category semantics: it considers the above two cases where both the meaning of aspect category and the sentiment polarity are removed.

The results are presented in the lower part in Ta-

	L14	R14	R15	R16
CMLA+ (Wang et al., 2017)	33.16	42.79	37.01	41.72
Li-unified-R (Li et al., 2019a)	42.34	51.00	47.82	44.31
P-pipeline (Peng et al., 2020)	42.87	51.46	52.32	54.21
Jet+BERT (Xu et al., 2020)	51.04	62.40	57.53	63.83
GTS+BERT (Wu et al., 2020)	55.21	64.81	54.88	66.08
Two-Stage (Huang et al., 2021)	<u>58.58</u>	68.16	58.59	67.52
GAS (Zhang et al., 2021)	58.19	<u>70.52</u>	<u>60.23</u>	<u>69.05</u>
PARAPHRASE	**61.13**	**72.03**	**62.56**	**71.70**

Table 3: Results of the ASTE task compared with previous state-of-the-art models. F1 scores are reported.

ble 2. We can see that discarding the semantics of either element leads to a performance drop, and the drop becomes larger after discarding both of them. Comparing the ablations on the sentiment polarity and aspect category, the model suffers more when the aspect category is projected to an indexed symbol. The possible reason is that there are only three types of sentiment polarities, which is much less than the number of types for the aspect category. Therefore, it can be easier for the model to learn the mapping between the special symbols and the polarity type during the training.

5.3 Results on ASTE and TASD Tasks

As described in Sec 3.4, the proposed PARAPHRASE modeling provides a unified framework to tackle the ABSA problem, we thus test it on the ASTE and TASD tasks, and compare with the previous state-of-the-art methods for each task.

For the ASTE task, we utilize the dataset provided by Xu et al. (2020). We adopt two types of baselines: 1) pipeline-based methods including **CMLA+** (Wang et al., 2017), **Li-unified-R** (Li et al., 2019a), **Peng-pipeline** (Peng et al., 2020) which firstly extract aspect and opinion terms separately, then conduct the pairing; **Two-stage** (Huang et al., 2021) which proposes a two-stage method to enhance the correlation between aspects and opinions; and 2) end-to-end models including **GTS** (Wu et al., 2020) and **Jet** (Xu et al., 2020), both designing unified tagging schemes in order to solve the task in an end-to-end fashion.

For the TASD task, we adopt the dataset prepared by Wan et al. (2020). We compare with a pipeline-type baseline method **Baseline-1-f_lex** (Brun and Nikoulina, 2018), two BERT based models including **TAS-CRF** and **TAS-TO** (Wan et al., 2020), and a recent model **MEJD** (Wu et al., 2021) which utilizes a graph structure to model the dependency among the sentiment elements.
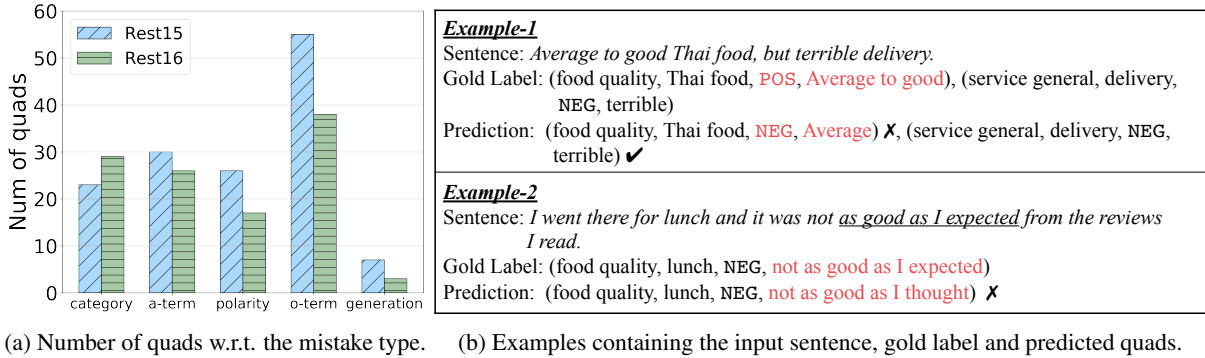
---

[3]The mapping relation between the category and their indexes is pre-defined based on the entire dataset.

(a) Number of quads w.r.t. the mistake type.

Example-1
Sentence: *Average to good Thai food, but terrible delivery.*
Gold Label: (food quality, Thai food, POS, Average to good), (service general, delivery, NEG, terrible)
Prediction: (food quality, Thai food, NEG, Average) ✗, (service general, delivery, NEG, terrible) ✔

Example-2
Sentence: *I went there for lunch and it was not as good as I expected from the reviews I read.*
Gold Label: (food quality, lunch, NEG, not as good as I expected)
Prediction: (food quality, lunch, NEG, not as good as I thought) ✗

(b) Examples containing the input sentence, gold label and predicted quads.

Figure 3: Error analysis and case study.

	Rest15	Rest16
Brun and Nikoulina (2018)	-	38.10
TAS-CRF (Wan et al., 2020)	57.51	65.89
TAS-TO (Wan et al., 2020)	58.09	65.44
MEJD (Wu et al., 2021)	57.76	67.66
GAS	60.63	68.31
PARAPHRASE	**63.06**	**71.97**

Table 4: Results of the TASD task compared with previous state-of-the-art models. F1 scores are reported.

The results for the ASTE and TASD tasks are shown in the Table 3 and 4 respectively. We also report the performance of the GAS method for comparison. We observe that the proposed PARA-PHRASE method consistently outperforms the previous state-of-the-art models across all datasets in two tasks, showing the effectiveness of converting various ABSA tasks into a paraphrase generation problem. More importantly, by transforming the problem into a unified S2S task, we alleviate extensive task-specific model designs. Unlike previous studies with different network architectures for different tasks, we use the same framework for solving the ASQP, ASTE, and TASD tasks, indicating the great generality of the PARAPHRASE method.

5.4 Error Analysis and Case Study

To better understand the behaviour of the proposed method, especially in which cases it would fail, we conduct error analysis and case study in this section. We sample 100 sentences in the development set of each dataset and employ the trained model to make the predictions. Then we check the incorrect quad predictions and categorize their error types.

We first analyze which type of sentiment element in the sentiment quad is the most difficult for the model to predict and present the results in Figure 3a. In both datasets, the most common mistake is

when predicting the opinion term. Different from the aspect term, opinion term is typically not a single word, but a text span. We find that the model often struggles to detect the exact same span as the ground-truths, as shown in the *Example-1* in Figure 3b. For the aspect category, the model is often confused by semantically similar aspect categories such as "*food quality*" and "*food style options*". For the sentiment polarity, the most common mistake is made by the confusion between "positive" and "neutral" classes, possibly due to the imbalanced label distribution in the dataset.

Moreover, we compute the amount of predicted quads whose sentiment elements do not belong to the corresponding vocabulary set, due to the nature of the generation modeling since it does not perform "extraction" in the given sentence. For instance, a predicted aspect category does not belong to the defined aspect category set V_c. As shown in the generation column in Figure 3a, this error type in fact accounts for only a small portion in total. *Example-2* presents a case for such error where the model changes the word "*expected*" in the original sentence to "*thought*" when predicting the opinion term. Although this might be similar to human readers, this prediction is judged as incorrect since we use the exact match for the evaluation. Nevertheless, contrary to the possible perception that the generation type method might generate unbounded contents which can be difficult to recover sentiment quads or provide meaningless outputs, the predictions from the proposed method actually suffer little from the generation error.

5.5 ABSA Cross-task Transfer

With the PARAPHRASE modeling, different ABSA tasks can be tackled in a similar manner, enabling the knowledge learned from related tasks to be easily transferred to the target task. In our case, ASTE
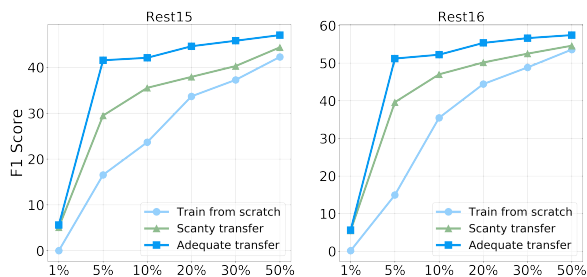
Figure 4: Cross-task transfer results. F1 scores on two datasets are shown with respect to the ratio of the ASQP data under three settings.

and TASD are regarded as two sub-tasks to transfer the knowledge for handling ASQP. Here we consider two common situations where we might have adequate ASTE/TASD data for transfer ("Adequate transfer") or we only have a small amount of ASTE/TASD data ("Scanty transfer"). In the experiment, we utilize 500/100 ASTE and TASD data samples for these two settings respectively. We vary the ratio of the ASQP data to simulate different scales of low-resource settings and report the results under two transfer situations in Figure 4. We also show the performance if we only train the model with the ASQP task, without any help from the knowledge transfer ("Train from scratch").

As can be observed in the figure, utilizing the knowledge learned from two triplet detection tasks can greatly benefit the concerned sentiment quad prediction. For instance, with adequate annotated data of ASTE and TASD, using 5% of the ASQP data can lead to competitive performance compared with purely training with 50% ASQP data. Even with a scanty amount of data from related tasks to transfer, the model can still perform much better than purely trained on the sentiment quad data, especially under the low-resource setting.

6 Conclusions

We introduce a new ABSA task, namely aspect sentiment quad prediction (ASQP) in this paper, aiming to provide a more comprehensive aspect-level sentiment picture. We propose a novel PARA-PHRASE modeling paradigm that tackles the original quad prediction as a paraphrase generation problem. Experiments on two datasets show its superiority compared with previous state-of-the-art models. We also demonstrate that the proposed method provides a unified framework that can be easily adapted to handle other ABSA tasks as well. Extensive analysis are conducted to understand the characteristics of the proposed method.

We can notice that ASQP remains a challenging problem and worth further exploring. We look forward future work could propose better methods to tackle such a difficult ABSA task for fully revealing the aspect-level opinion information.

References

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020. Augmented natural language for generative sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 375–385, Online. Association for Computational Linguistics.

Rahul Bhagat and Eduard H. Hovy. 2013. What is a paraphrase? *Comput. Linguistics*, 39(3):463–472.

Caroline Brun and Vassilina Nikoulina. 2018. Aspect based sentiment analysis into the wild. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2018*, pages 116–122.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A chinese review dataset towards aspect category sentiment analysis and rating prediction. *CoRR*, abs/2103.06605.

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*, pages 833–843.

Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 6515–6524.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *AAAI*.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 3685–3694.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

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 *ACL19*, pages 504–515.

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2019a. CAN: constrained attention networks for multi-aspect sentiment analysis. 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*, pages 4600–4609.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019b. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 537–546.

Binxuan Huang and Kathleen M. Carley. 2018. Parameterized convolutional neural networks for aspect level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1091–1096.

Lianzhe Huang, Peiyi Wang, Sujian Li, Tianyu Liu, Xiaodong Zhang, Zhicong Cheng, Dawei Yin, and Houfeng Wang. 2021. First target and opinion then polarity: Enhancing target-opinion correlation for aspect sentiment triplet extraction. *CoRR*, abs/2102.08549.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *AAAI*, pages 6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP 2019*, pages 34–41.

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

Pengfei Liu, Shafiq R. Joty, and Helen M. Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 1433–1443.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *CoRR*, abs/2103.10385.

Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. 2019. DOER: dual cross-shared RNN for aspect term-polarity co-extraction. In *ACL*, pages 591–601.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 3538–3547.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. *CoRR*, abs/2101.00816.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

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 *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia 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*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *SemEval@NAACL-HLT*, 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 *SemEval@COLING 2014*, pages 27–35.

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

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 999–1005.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, NAACL-HLT 2019, pages 380–385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008.

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*, pages 9122–9129.

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*, pages 3316–3322.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 606–615.

Chao Wu, Qingyu Xiong, Hualing Yi, Yang Yu, Qiwu Zhu, Min Gao, and Jie Chen. 2021. Multiple-element joint detection for aspect-based sentiment analysis. *Knowledge-Based Systems*, page 107073.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pages 592–598.

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*, pages 2339–2349.

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*, pages 2416–2429.

Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*, pages 2979–2985.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 3540–3549.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. 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*, pages 504–510.

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

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015. Representation learning for aspect category detection in online reviews. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 417–424.

9219