Question-Driven Span Labeling Model for Aspect–Opinion Pair Extraction

Lei Gao,[1,3] Yulong Wang,[1,3*] Tongcun Liu,[2,3*] Jingyu Wang,[1,3] Lei Zhang,[1,3] Jianxin Liao[1,3]

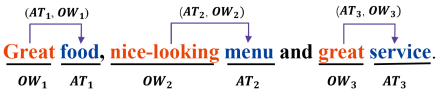
[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[2]School of Information Engineering, Zhejiang A & F University
[3]EBUPT Information Technology Co., Ltd.
{gaolei_1, wangyulong, liutongcun, wangjingyu, zhanglei, liaojianxin}@ebupt.com

Abstract

Aspect term extraction and opinion word extraction are two fundamental subtasks of aspect-based sentiment analysis. The internal relationship between aspect terms and opinion words is typically ignored, and information for the decision-making of buyers and sellers is insufficient. In this paper, we explore an aspect–opinion pair extraction (AOPE) task and propose a Question-Driven Span Labeling (QDSL) model to extract all the aspect–opinion pairs from user-generated reviews. Specifically, we divide the AOPE task into aspect term extraction (ATE) and aspect-specified opinion extraction (ASOE) subtasks; we first extract all the candidate aspect terms and then the corresponding opinion words given the aspect term. Unlike existing approaches that use the BIO-based tagging scheme for extraction, the QDSL model adopts a span-based tagging scheme and builds a question–answer-based machine-reading comprehension task for an effective aspect–opinion pair extraction. Extensive experiments conducted on three tasks (ATE, ASOE, and AOPE) on four benchmark datasets demonstrate that the proposed method significantly outperforms state-of-the-art approaches.

Introduction

An aspect-based sentiment analysis (ABSA) task involves identifying opinions expressed toward specific entities, e.g., the price of a laptop (Li and Lam 2017). This task involves two closely related subtasks: Aspect term extraction (ATE) and opinion word extraction (OWE). As a fundamental subtask of an ABSA, the objective of the ATE is to extract the aspect term (i.e., a word or phrase) that describes an attribute or feature of an entity in a given sentence (Pontiki et al. 2014, 2015, 2016). The purpose of the OWE is to extract opinion words, which are expressions carrying subjective emotions in a sentence (Liu, Xu, and Zhao 2014). Earlier works focused only on the ATE task (Jakob and Gurevych 2010; Li et al. 2010; Liu, Xu, and Zhao 2012; Mukherjee and Liu 2012), and did not consider the internal relationship between ATE and OWE subtasks. None of these works could provide sufficient information for the decision-making of buyers and sellers.



(a).ATE task: { food, menu, service }

(b).OWE task: { Great, nice-looking, great }

(c).AOPE task: { <food, Great>, <menu, nice-looking>, <service, great> }

Figure 1: Example showing the differences between ATE, OWE, and AOPE tasks.

To solve these issues, the potential opinion information in a sentence has been utilized to improve the performance of the ATE task (Liu, Xu, and Zhao 2012; Liu et al. 2013; Li and Lam 2017). Moreover, studies have shown that using the information of the aspect terms and opinion words in a sentence can help to mutually improve the performance of ATE and OWE tasks (Wang and Wang 2008; Qiu et al. 2011; Liu et al. 2013; Yu, Jiang, and Xia 2019). Although these approaches outperform conventional ones, the objective is to extract the aspect term set and/or opinion word set from a given sentence, rather than extracting the aspect–opinion pairs, which have great significance in reality.

Therefore, in this study, we focus on an aspect–opinion pair extraction (AOPE) subtask for ABSA, which aims to extract all the aspect–opinion pairs from a review text. For instance, in the sentence "Great food, nice-looking menu and great service." the words "food", "menu" and "service" are aspect terms, whereas the words "great", "nice-looking" and "great" are their corresponding opinions; these can be extracted by the ATE and OWE tasks, respectively. However, the AOPE task involves extracting the aspect–opinion pairs set {<food, great>, <menu, nice-looking>, <service, great>}. Figure 1 illustrates the detailed difference between the ATE, OWE, and AOPE tasks.

The aspect–opinion pairs can ensure a more fine-grained sentiment analysis for review texts and will benefit many downstream applications such as opinion summarization and product profiling (Zhao et al. 2020). By referring to the aspect–opinion pairs in a review sentence, customers can rapidly obtain a glimpse of the pros and cons of a product or service. Despite the significance of ABSA, few works have studied AOPE tasks, owing to the following challenges: (1) Pairing problems: for different aspect terms, the corre-

[*]Corresponding author and contributing equally with the first author.

Pairing problem	*For different aspect terms, the corresponding opinion words may be different:* The food is great (big selection , reasonable prices) and the drinks are really good . {<food, great)>, (selection, big)>, (prices, reasonable)>, (drinks, really good)>}
Overlapping problem	*Same aspect term, opinion words overlapping*: I highly recommend the grand marnier shrimp , it 's insanely good . {<grand marnier shrimp, recommend>, <grand marnier shrimp, insanely good>}
Missing values problem	*There are some aspect terms in the sentence that do not have their corresponding opinion words*: The price is so cheap , but that does not reflect the service or the atmosphere . {<price, cheap>}
Nesting problem	*Some words like "fresh" and "artificial" may belong to both aspect terms and opinion words.* They used artificial lobster meat but their service is prompt and accurate . {<artificial lobster meat, artificial>, <service, prompt>, <service, accurate>}

Figure 2: Examples illustrating the challenges of the AOPE task.

sponding opinion words may be different, and vice versa; (2) Overlapping problems: different aspect–opinion pairs may overlap in a sentence; (3) Missing values problems: not all aspect terms in a sentence have corresponding opinion words; (4) Nesting problems: in reality, some words in a review text may be labeled as part of both aspect terms and opinion words (Yu, Jiang, and Xia 2019). Figure 2 shows some examples of these challenges. For the missing values problem, the blue underlined words indicate aspect terms with no corresponding opinion words in the review text; for the nesting problem, the words in green may be either an aspect term or an opinion word.

To solve the above-mentioned challenges, we designed a Question-Driven Span Labeling (QDSL) model to extract aspect–opinion pairs from review texts. Our method decomposes the AOPE task into two subtasks: aspect term extraction (ATE) and aspect-specified opinion extraction (ASOE). Specifically, we first extract all the candidate aspect terms from the comment text; subsequently, we construct an auxiliary question for each candidate aspect term; finally, we combine the special question constructed for each candidate aspect with the original review text as sentence pairs, and naturally formulate the ASOE task as a machine-reading comprehension (MRC) task. To sum up, we first extract all the candidate aspect terms and then the corresponding opinion words for each aspect term to solve the AOPE task. All the aforementioned challenges in the AOPE task can be resolved with this design. In summary, the main contributions of this paper are as follows:

- We explore an AOPE task and propose a QDSL model, which divides the AOPE task into ATE and ASOE subtasks. By this design, the negative effects of error propagation and redundancy pairs faced by extract-then-classify methods can be alleviated.
- We adopt a span-based rather than BIO-based tagging scheme for extraction. The span-based tagging scheme can be applied to complex situations where a token belongs to multiple different entities.
- We formulate the ASOE task as an MRC problem rather than a sequence labeling problem. By solving the ASOE task from this perspective, the QDSL model can better capture the aspect-specified prior features and has excellent interpretability.
- Extensive experiments were conducted on four datasets, and the results show that our QDSL model can yield significant performance improvement over state-of-the-art approaches.

Related Work

The ATE has been extensively studied. Conventional methods for ATE tasks can be divided into unsupervised (Liu, Xu, and Zhao 2012), semi-supervised (Mukherjee and Liu 2012), and supervised (Jakob and Gurevych 2010; Li et al. 2010) methods. Recently, deep neural network-based methods have shown significant performance improvement for ATE tasks. For example, Xu et al. (2018) used a double embedding-based CNN model to extract aspect terms. Unlike most deep learning methods that treat ATE tasks as sequence labeling tasks, Ma et al. (2019) and Li et al. (2020a) explored a seq2seq framework for ATE tasks. However, these works extracted aspect terms without considering the information of the opinion words in the sentences.

Recently, a new research direction, which aims at co-extracting the aspect and opinion terms, has drawn increasing attention in both academia and industry (Wang and Wang 2008; Qiu et al. 2011; Liu et al. 2013; Yu, Jiang, and Xia 2019). Some methods have achieved significant progress on both subtasks. For example, Qiu et al. (2011) designed a bootstrapping-based double-propagation mechanism to expand the initial opinion lexicon and extract targets. Liu et al. (2013) utilize a partially-supervised word alignment model to extract aspect terms and opinion words jointly. Recently, other works have used a deep neural network-based multi-task learning framework to jointly extract aspect terms and opinion words from review texts (Wang et al. 2016, 2017; Wang and Pan 2018, 2019; Yu, Jiang, and Xia 2019). These approaches have outperformed conventional ones; however, none of these works considered the aspect terms and opinion words as pairs.

Only a few works have studied AOPE tasks. For example, Klinger and Cimiano (2013a,b) and Yang and Cardie (2013) explored joint learning models for AOPE tasks. However,

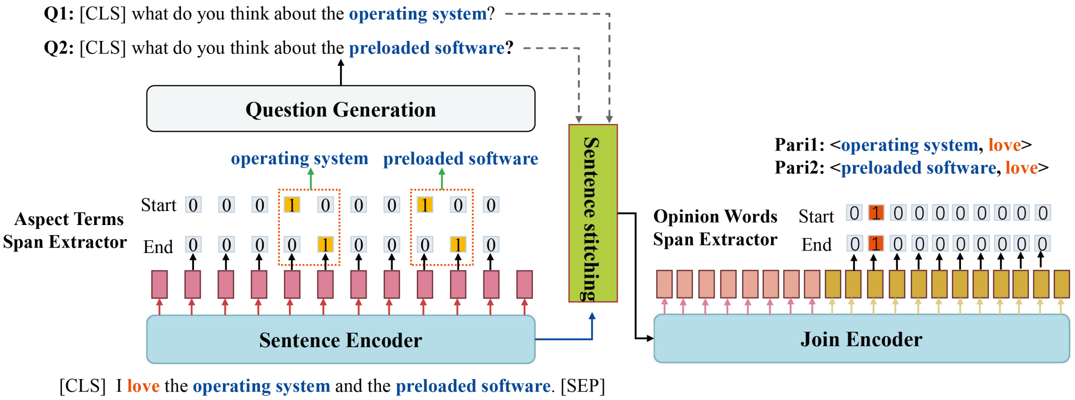

Q1: [CLS] what do you think about the operating system?
Q2: [CLS] what do you think about the preloaded software?
Question Generation
operating system
preloaded software
Aspect Terms Span Extractor
Start 0 0 0 1 0 0 0 1 0 0
End 0 0 0 0 1 0 0 0 1 0
Sentence Encoder
[CLS] I love the operating system and the preloaded software. [SEP]
Sentence stitching
Pari1: <operating system, love>
Pari2: <preloaded software, love>
Opinion Words Span Extractor
Start 0 1 0 0 0 0 0 0 0 0
End 0 1 0 0 0 0 0 0 0 0
Join Encoder

Figure 3: Illustration of the proposed QDSL model.

these methods rely heavily on external syntax resources and hand-crafted features. Recently, Zhao et al. (2020) developed a span-based multi-task learning framework for AOPE tasks. Chen et al. (2020) proposed a synchronous double-channel recurrent network (SDRN) to integrate high-level interaction information for AOPE.

Unlike current works, our method use a deep neural network to extract features automatically, without reliance on external language knowledge and dictionary. Compared with the current extract-then-classify methods (Zhao et al. 2020; Chen et al. 2020), our method yields better performance and interpretability.

Methodology

Problem Definition

Given an input sentence $\mathcal{S} = \{w_1, w_2, \ldots, w_N\}$ with N words, the objective of the AOPE task is to extract a set of all <aspect, opinion> pairs $\mathcal{P} = \{< at_1, ow_1 >, \ldots, < at_m, ow_m >\}$ from a sentence. Note that at_m and ow_m could be a single word or a phrase. Inspired by previous works (Wei et al. 2020; Sun, Huang, and Qiu 2019; Li et al. 2019), we process this task through a span-based extraction approach in a question-driven manner. Formally, we decompose the AOPE task into two subtasks: ATE and ASOE. For the ATE subtask, the objective is to extract a collection of all the explicitly mentioned aspect terms $\mathcal{AT} = \{at_1, at_2, \ldots, at_{|\mathcal{AT}|}\}$ from the review sentence S. For the ASOE subtask, which is sometimes called target-oriented opinion word extraction (TOWE) (Fan et al. 2019; Wu et al. 2020), the objective is to identify and extract a collection of all the corresponding opinion words $\mathcal{OW} = \{ow_1, ow_2, \ldots, ow_{|\mathcal{OW}|}\}$ for a given aspect term from the review sentence. Finally, by combining the two subtasks, we can extract all the aspect–opinion pairs in the sentence $\mathcal{S}$ simultaneously.

Model Description

Unlike previous works that typically extracted a set of aspect terms and another set of opinion words separately, our objective is to extract all the aspect–opinion pairs from a given sentence. Figure 3 shows the architecture of the proposed QDSL model, which comprise two extractors: the aspect terms span extractor (ATSE) and opinion words span extractor (OWSE). The details of each component of the QDSL model are given in the following sections.

Encoder

BERT is the most commonly employed encoder for extracting context-sensitive features for downstream tasks (Devlin et al. 2019). In our model, we use BERT as a sentence encoder and joint encoder. In the sentence encoder part, BERT was employed to encode context-related features for a given sentence; in the joint encoder part, BERT was utilized to encode the constructed auxiliary question and the original sentence pair. Through the bidirectional self-attention mechanism in BERT, the prior information of the aspect term can be encoded into the representation of the original sentence for subsequent ASOE subtask.

Aspect Terms Span Extractor

We utilized BERT to generate bidirectional representations $\mathbf{H}^A = \{\mathbf{h}_1^A, \mathbf{h}_2^A, \ldots, \mathbf{h}_N^A\}$ of the original review $\mathcal{S}$, where the superscript A indicates that the extracted feature was used for the ATE subtask.

Previous works mostly formalized ATE as a sequence labeling problem based on the BIO (Xu et al. 2018) or BMES (Yin et al. 2016) tagging schemes. However, these sequence-labeling models can only assign one label to each token and are unsuitable for complex situations where a token may belong to multiple entities (Li et al. 2020b). Inspired by the recent advances in relation extraction (Wei et al. 2020), we designed an ATSE to extract all the candidate aspect terms in a review sentence. Specifically, we adopted a span-based scheme rather than a BIO-based scheme to extract the aspect terms, i.e., two binary classifiers are used to detect the start and end positions of each aspect item:

$$p_i^{at-s} = \sigma(\mathbf{w}^{at-s}\mathbf{h}_i^A + b^{at-s}) \tag{1}$$

$$p_i^{at-e} = \sigma(\mathbf{w}^{at-e}\mathbf{h}_i^A + b^{at-e}) \tag{2}$$

where p_i^{at-s} and p_i^{at-e} represent the probability that the i-th word is at the beginning or at the end of an aspect term, respectively. The predicted results $\hat{y}_i^{at-s}$ and $\hat{y}_i^{at-e}$ of the ATSE task were generated from the predicted probability distributions of p_i^{at-s} and p_i^{at-e}:

$$\hat{y}_i^{at-s} = \begin{cases} 1, & \text{if } p_i^{at-s} > 1 - p_i^{at-s}; \\ 0, & \text{else.} \end{cases} \tag{3}$$

$$\hat{y}_i^{at-e} = \begin{cases} 1, & \text{if } p_i^{at-e} > 1 - p_i^{at-e}; \\ 0, & \text{else.} \end{cases} \tag{4}$$

Note that we match each $\hat{y}_i^{at-s} = 1$ with its nearest $\hat{y}_i^{at-e} = 1$ as a start–end pair to determine all the aspect terms in the sentence.

The loss function for the aspect term extraction subtask can be formulated using the binary cross-entropy error between the predicted and gold scores:

$$\mathcal{L}_{\text{ATSE}} = \mathcal{L}_{\text{ATSE}}^{s} + \mathcal{L}_{\text{ATSE}}^{e} \tag{5}$$

$$= \sum_{i=1}^{N} \sum_{at \in \{at-s, at-e\}} \text{BCE}(p_i^{at}, y_i^{at}) \tag{6}$$

where BCE is the binary cross entropy function.

Question Generator

This component plays an important role in our framework to undertake ATSE and OWSE extractors. We must transfer the extraction results of the ATSE to the OWSE component as prior information to solve the ASOE subtask. Naturally, we decided to formulate the ASOE subtask as an MRC problem rather than a sequence-labeling problem. Therefore, the objective of the question generator is to generate appropriate question sentences for the ASOE component. Inspired by a previous work (He, Lewis, and Zettlemoyer 2015), for simplicity, we generate one question for each aspect term at_m using a fixed template "What do you think about the at_m?" For example, given a sentence "I love the operating system and the preloaded software," the aspect terms are "operating system" and "preloaded software"; thus, the generated questions are "what do you think about the operating system?" and "what do you think about the preloaded software?"

Opinion Words Span Extractor

The OWSE is designed to extract all the opinion words given the extracted candidate aspect terms and the original sentence. Therefore, the OWSE should have the ability to generate aspect-specified features for extraction when given different aspect terms. In this section, we first introduce the standard OWSE approach and then present the proposed question-driven OWSE.

Standard OWSE For a given aspect term at_m and the original review S, we assume that $at_m = \{w_k, \dots, w_l\}$. The standard OWSE (Stand-OWSE) can extract the aspect-specified sentence features via the following equations:

$$\mathbf{H}^O = \{\mathbf{h}_1^O, \mathbf{h}_2^O, \dots, \mathbf{h}_n^O\} \tag{7}$$

$$\mathbf{h}_i^O = \mathbf{h}_i^A + \mathbf{v}^A \tag{8}$$

$$\mathbf{v}^A = \frac{\sum_{i=k}^{l} \mathbf{h}_i^A}{l - k} \tag{9}$$

where the superscript O indicates the extracted feature from BERT for the ASOE subtask. In the Stand-OWSE, $\mathbf{h}_i^A$ can be obtained directly from the output of the ATSE component.

Question-Driven OWSE Unlike the standard OWSE, we combine the generated auxiliary question sentence from the question generator and the original sentence to generate a sentence pair <Question, Answer>. For the example shown in Figure 3, the new constructed sentence is "[CLS] what do you think about the operating system? [SEP] I love the operating system and preloaded software. [SEP]." Subsequently, the new sentence is encoded by BERT. Through the bidirectional cross attention between the two sentences, the prior information of the aspect term can be directly encoded into the representation of the original sentence. Thus, we can obtain the representation $\mathbf{H}^O$ of the original sentence for the ASOE subtask.

Similar to the ATSE, we still use span-based scheme to extract the aspect-specified opinion words. Here, we define the loss of the aspect-specified opinion extraction as $\mathcal{L}_{\text{OTSE}}^{at_m}$.

Joint Learning

Finally, the overall loss function can be expressed as follows:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{ATSE}} + (1 - \lambda) \sum_{at_m \in \mathcal{AT}} \mathcal{L}_{\text{OTSE}}^{at_m} \tag{10}$$

where $\mathcal{AT}$ is the collection of all aspect terms, λ is a hyperparameter.

Experiments

Datasets

We evaluated the performance of our proposed QDSL model on four public datasets obtained from SemEval 2014 Task 4, SemEval 2015 Task 12, and SemEval 2016 Task 5. These datasets are widely used in ABSA tasks. We use $\mathbb{S}_{14l}$, $\mathbb{S}_{14r}$, $\mathbb{S}_{15r}$, and $\mathbb{S}_{16r}$ to denote SemEval-2014 Laptops, SemEval-2014 Restaurants, SemEval-2015 Restaurants, and SemEval-2016 Restaurants datasets, respectively. These datasets come from the SemEval challenge, in which only aspect term annotations are provided. Therefore, Fan et al. (2019) annotated the SemEval dataset with the corresponding opinion words for each aspect term.

To ensure a comprehensive comparison, we designed three different comparative experiments for the ATE, ASOE, and AOPE tasks. Therefore, for the experiment on the ATE subtask, we used the original SemEval datasets to compare our method with the other methods and keep the official data division of these datasets for the training, validation, and testing sets. Table 1 lists the statistics of the original dataset. For the experiment on the OWE and AOPE tasks, we only used the datasets provided by Fan et al. (2019). Similar to previous works (Fan et al. 2019; Wu et al. 2020), we randomly split 20% of the training set as the validation set. Table 2 lists the statistics of these datasets, where approximately 28.5% of the sentences have overlapping aspect terms or opinion words.

Datasets	$\mathbb{S}_{14l}$			$\mathbb{S}_{14r}$			$\mathbb{S}_{15r}$			$\mathbb{S}_{16r}$		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
#Sen	3048	100	800	3044	100	800	967	36	466	1429	37	478
#Aspect	2325	49	651	3664	96	1130	1110	38	496	1609	49	562
#Sen w/ aspect terms	1492	40	422	2023	54	606	780	30	366	1147	31	390

Table 1: Statistics of the original SemEval datasets obtained from Pontiki et al. (2014, 2015, 2016).

Datasets	$\mathbb{S}_{14l}$		$\mathbb{S}_{14r}$		$\mathbb{S}_{15r}$		$\mathbb{S}_{16r}$	
	Train	Test	Train	Test	Train	Test	Train	Test
#Sentences	1151	343	1625	500	754	325	1079	328
#Aspect-opinion term pairs	1784	535	2892	939	1228	482	1710	514
#Sentences w/ overlapping	334	100	517	189	210	70	280	82

Table 2: Statistics of the datasets provided by Fan et al. (2019).

Experimental Settings

The pre-trained BERT model used in the QDSL model is $\text{BERT}_{\text{BASE-UNCASED}}$, which is trained on lowercased English text. Based on the fine-tuning hyperparameters suggested by Devlin et al. (2019), the learning rate was set to 5e-5, the batch size was set to 16, and the dropout probability was set to 0.1. We used AdamW (Loshchilov and Hutter 2019) to optimize the model parameters. In the experiment, we trained a total of 40 epochs and selected the best performing model in the validation set for testing.

Compared Approaches

We conducted experiments on ATE, ASOE, and AOPE tasks, and compared our method with the following state-of-the-art approaches.

Aspect Term Extraction We compare the ATSE component in the QDSL model with the following approaches for the ATE task on the original SemEval dataset.

- **CMLA** (Wang et al. 2017): CMLA is a multilayer attention network that extracts aspect terms and opinion words separately.
- **HAST** (Li et al. 2018): HAST exploits two useful clues, namely opinion summary and aspect detection history, to achieve more accurate aspect term extraction.
- **Seq2Seq4ATE** (Ma et al. 2019): For Seq2Seq4ATE, the ATE task is formalized as a sequence-to-sequence (Seq2Seq) learning task.
- **BERT-ATE** (Xu et al. 2019): BERT-ATE is a BERT-based neural network model for the ATE task. In addition, this method uses the BIO-based tagging scheme for extraction.

Aspect Specified Opinion Extraction We compared the OWSE component in the QDSL model with the following state-of-the-art approaches for the ASOE task.

- **IOG-{Greedy,CRF}** (Fan et al. 2019): IOG divides a sentence into three parts based on the given aspect term, and uses Bi-LSTM to capture the aspect-specified information in these three parts. Finally, two different decoding methods (Greedy decoding, CRF) were used to extract the target-oriented opinion words.
- **LOTN** (Wu et al. 2020): LOTN uses the latent opinion information in other sentiment classification datasets to improve the performance of TOWE tasks.

Aspect–Opinion Pair Extraction We compared our QDSL model with the following baselines to validate the efficiency of our model for the AOPE task.

- **HAST+IOG**: This is a pipeline method that combines HAST (Li et al. 2018) and IOG (Fan et al. 2019).
- **JERE-MHS** (Bekoulis et al. 2018): JERE-MHS is a model for joint entity-relation extraction. This model can simultaneously detect both entity types and relationship types. Therefore, this method can be applied to AOPE tasks.
- **SDRN** (Chen et al. 2020): SDRN is composed of an opinion entity extraction unit, relationship extraction unit and a synchronous unit to extract aspect-opinion pairs.
- **SpanMlt** (Zhao et al. 2020): SpanMlt formulates the AOPE task as a joint term and relation extraction problem and develops a multi-task learning framework.

Evaluation Metrics

Following the evaluation metrics used in the previous work (Chen et al. 2020; Zhao et al. 2020), we use the F_1 score metric to evaluate the performance of our model and the compared approaches for the ATE, ASOE, and AOPE subtasks. Note that an aspect–opinion pair is considered correct if and only if both the aspect term and the corresponding opinion word are predicted correctly.

Results and Discussion

We first conducted an experiment on an AOPE task to evaluate the performance of our QDSL model. We then conducted experiments on ATE and ASOE tasks to prove that

Methods	Datasets			
	$\mathbb{S}_{14l}$	$\mathbb{S}_{14r}$	$\mathbb{S}_{15r}$	$\mathbb{S}_{16r}$
HAST+IOG	53.41	62.39	58.12	63.84
JERE-MHS	53.34	66.02	59.64	67.65
SDRN	67.13	76.48	70.94	-
SpanMlt	68.66	75.60	64.48	71.78
Stand-SL	64.37	61.54	61.15	66.87
QDSL	**70.20**	**78.05**	**71.22**	**77.28**

Table 3: F_1 scores for the AOPE task on four datasets.

Methods	Datasets			
	$\mathbb{S}_{14l}$	$\mathbb{S}_{14r}$	$\mathbb{S}_{15r}$	$\mathbb{S}_{16r}$
CMLA	77.80	85.29	70.43	72.77
HAST	79.52	85.61	71.46	73.61
Seq2seq4ATE	80.31	-	-	75.14
BERT-ATE	79.28	-	-	74.10
ATSE	**84.27**	**87.85**	**77.72**	**83.34**

Table 4: F_1 scores for the ATE task on four datasets.

the ATSE and question-driven OWSE can significantly improve the performance.

Results for AOPE Task The comparison results of the AOPE task are reported in Table 3. According to the results, our QDSL model achieved the highest F_1 score for the AOPE task on the four datasets. The performance of the Stand-SL model was lower than those of SpanMlt and QDSL. This occurred because Stand-SL uses the standard opinion word span extraction method and cannot accurately extract the aspect-specified features required for the AOPE task.

Figure 4 shows prediction results generated by the Stand-SL and QDSL for some examples. As shown in the first and fourth rows, Stand-SL cannot extract the aspect-specified information effectively, so the model attempts to extract the same opinion words for different aspect terms. The QDSL can accurately extract the required prior information for the ASOE subtask. For the third and fourth cases, although Stand-SL and QDSL have the same aspect term extraction component, Stand-SL cannot extract all aspect terms accurately, but QDSL does. This result indicates that QDSL can better combine ATE and ASOE subtasks so that the two components can benefit from each other. In addition, both models can understand the complex semantic information hidden in the sentences. For example, in the second sentence, both models extract actual aspect opinion pairs that do not even appear in the golden annotation. Evidently, QDSL performs better. It even extracts nested aspect terms and opinion words: “resolution” is not only a part of opinion words, but is also an aspect term.

Result for ATE Task We compared the performance of the ATSE component, namely ATSE, with those of the state-of-the-art approaches for the ATE task. Table 4 lists the results. The results show that our model significantly outperforms existing methods. Because Seq2seq4ATE and BERT-ATE are both methods based on BERT, the results also indicate that the improvement achieved by ATSE is attributable not only to BERT but also to the tagging scheme used in the decoding stage. For the BIO-based tagging scheme, the search space of the decoder will increase exponentially with the sequence length, which will increase the difficulty of modeling the relationship between tags. The span-based tagging scheme used in our QDSL model can not only extract the aspect term intuitively but can also reduce the parameter space, making the model easy to optimize. Moreover, the span-based approach can identify overlapping entities such as “service” within “table service.”

Methods	Datasets			
	$\mathbb{S}_{14l}$	$\mathbb{S}_{14r}$	$\mathbb{S}_{15r}$	$\mathbb{S}_{16r}$
IOG-Greedy	71.35	80.02	73.25	81.69
IOG-CRF	71.39	80.24	73.51	81.84
LOTN	72.02	82.21	73.29	83.62
S-OWSE	64.98	57.81	65.98	73.31
QD-OWSE	**80.35**	**87.23**	**80.71**	**88.14**

Table 5: F_1 scores for the ASOE task on four datasets.

Results for ASOE Task Table 5 lists the results for the ASOE task. S-OWSE is the variant model that uses Standard OWSE, whereas QD-OWSE is the variant model that uses the question-driven OWSE. The results show that LOTN as the state-of-the-art approach outperforms IOG, and that performance of S-OWSE is lower than that of LOTN. However, the QD-OWSE proposed in this paper yields a higher F_1 score than LOTN. This shows that the proposed question-driven method can accurately extract the aspect-specified features required to perform ASOE subtasks.

Model Analysis

Base Encoder We further explored the effectiveness of the different BERT encoders for our framework. The results are listed in Table 6. QDSL$_{\text{RANDOM}}$ is the QDSL framework in which the BERT parameters are randomly initialized. QDSL$_{\text{FIXED}}$ is the QDSL framework in which the parameters of the BERT embedding layers are fixed. QDSL$_{\text{FINE-TUNE}}$ is the normal QDSL.

The results indicat that the pre-trained language model is very important for BERT-based downstream tasks. Directly performing AOPE tasks on the BERT model without pre-training does not yield good performance. In our model, the parameters for sentence encoder and joinit encoder were shared; this enable the internal relationship between the closely related ATE and ASOE subtasks can be better modeled, resulting in performance exceeding that of the single-task model on both tasks.

Multi-task Setup Because our QDSL is a multi-task learning framework, the loss function was composed of two parts. Thus, we further investigated the balance between the two subtasks for multi-task learning. We set different loss proportions to train the model. Here, λ was varied between

Input	Challenges	Golden	Stand-SL	QDSL
It is a great size and amazing windows 8 included!	Pairing problem	<size, great>, <windows 8, amazing>	<size, great>, ✓ <size, amazing>, ✗ <windows 8, great>, ✗ <windows 8, amazing> ✓	<size, great>, ✓ <windows 8, amazing> ✓
Pizzas were excellent in addition to appetizers and main courses.	Pairing & Overlapping problem	<Pizzas, excellent>	<Pizzas, excellent>, ✓ <appetizers, excellent>, ✯ <main courses, excellent> ✯	<Pizzas, excellent>, ✓ <appetizers, excellent>, ✯ <main courses, excellent> ✯
the screen , the software and the smoothness of the operating system	Pairing & Missing values problem	<operating system, smoothness>	<screen, smoothness>, ✗ <software, smoothness>, ✗ <operating system, smoothness> ✓	<operating system, smoothness> ✓
air has higher resolution but the fonts are small.	Pairing & Nesting problem	<resolution, higher>, <fonts, small>	<air, higher>, ✗ <air, small >, ✗ <resolution, higher>, ✓ <resolution, small> ✗	<air, higher resolution>, ✯ <resolution, higher>, ✓ <fonts, small> ✓

Figure 4: Examples predicted by Stand-SL and QDSL. ✓ indicates a correct prediction, ✗ indicates an incorrect prediction, and ✯ indicates a correct prediction not in the golden annotation.

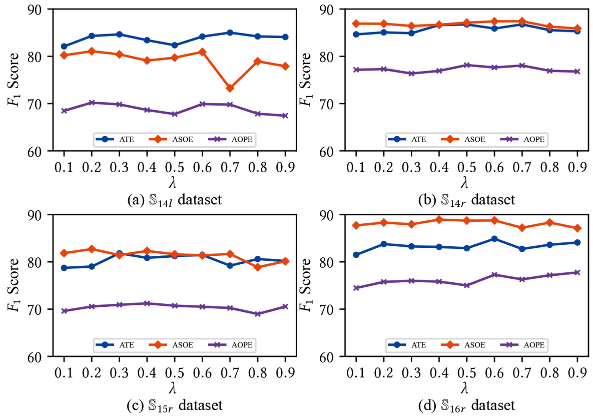
Datasets	$\mathbb{S}_{14l}$			$\mathbb{S}_{14r}$			$\mathbb{S}_{15r}$			$\mathbb{S}_{16r}$		
	ATE	ASOE	AOPE	ATE	ASOE	AOPE	ATE	ASOE	AOPE	ATE	ASOE	AOPE
ATSE	84.27	-	-	87.85	-	-	77.72	-	-	83.34	-	-
QD-OWSE	-	80.35	-	-	87.23	-	-	80.71	-	-	88.14	-
QDSL$_{\text{RANDOM}}$	52.57	48.11	29.47	62.40	50.68	35.48	54.05	48.53	32.83	57.11	52.52	33.67
QDSL$_{\text{FIXED}}$	83.18	79.13	67.91	86.33	86.38	76.85	80.82	82.84	71.41	83.35	89.58	76.86
QDSL$_{\text{FINE-TUNE}}$	**84.31**	**81.07**	**70.20**	**86.75**	**87.41**	**78.05**	**80.86**	**82.32**	**71.22**	**84.89**	**88.78**	**77.28**

Table 6: Results of QDSL with different base encoders.

0.1 and 0.9, and increased by 0.1. Figure 5 shows the results. Evidently, the performance of the model on these four datasets is very stable, and the F1 score does not change significantly with the change in λ. The experimental results show that the QDSL model is robust and insensitive to the hyperparameters.

Conclusions

In this study, we explored the AOPE task and designed a QDSL model. Unlike existing extract-then-classify methods, we divided the AOPE task into ATE and ASOE subtasks. Through this design, we overcame the drawbacks of the current extract-then-classify methods for AOPE tasks, and focused on designing an effective model for the ASOE subtask to capture the aspect-specified information. Intuitively, we guide the ASOE subtask through a question to formulate the ASOE subtask as an MRC task. This design is extremely reasonable and explainable. Extensive experiments on four datasets showed that the proposed method significantly outperforms state-of-the-art models for AOPE tasks. Our QDSL model can be used not only for aspect-based sentiment analysis but also for aspect-level sentiment classification (ASC) and aspect category detection (ACD). In the future, we will apply our QDSL to these tasks.



(a) $\mathbb{S}_{14l}$ dataset

(b) $\mathbb{S}_{14r}$ dataset

(c) $\mathbb{S}_{15r}$ dataset

(d) $\mathbb{S}_{16r}$ dataset

Figure 5: F_1 curves on four datasets for the three tasks using the best model setup when adjusting the loss balance.

Acknowledgements

This work was supported in part by the National Key R&D Program of China 2020YFB1807800, and in part by the National Natural Science Foundation of China under Grants 61771068.

References

Bekoulis, G.; Deleu, J.; Demeester, T.; and Develder, C. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.* 114: 34–45.

Chen, S.; Liu, J.; Wang, Y.; Zhang, W.; and Chi, Z. 2020. Synchronous Double-channel Recurrent Network for Aspect-Opinion Pair Extraction. In *ACL*, 6515–6524. Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

Fan, Z.; Wu, Z.; Dai, X.; Huang, S.; and Chen, J. 2019. Target-oriented Opinion Words Extraction with Target-fused Neural Sequence Labeling. In *NAACL-HLT (1)*, 2509–2518. Association for Computational Linguistics.

He, L.; Lewis, M.; and Zettlemoyer, L. 2015. Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language. In *EMNLP*, 643–653. The Association for Computational Linguistics.

Jakob, N.; and Gurevych, I. 2010. Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields. In *EMNLP*, 1035–1045. ACL.

Klinger, R.; and Cimiano, P. 2013a. Bi-directional Interdependencies of Subjective Expressions and Targets and their Value for a Joint Model. In *ACL (2)*, 848–854. The Association for Computer Linguistics.

Klinger, R.; and Cimiano, P. 2013b. Joint and Pipeline Probabilistic Models for Fine-Grained Sentiment Analysis: Extracting Aspects, Subjective Phrases and their Relations. In *ICDM Workshops*, 937–944. IEEE Computer Society.

Li, F.; Han, C.; Huang, M.; Zhu, X.; Xia, Y.; Zhang, S.; and Yu, H. 2010. Structure-Aware Review Mining and Summarization. In *COLING*, 653–661. Tsinghua University Press.

Li, K.; Chen, C.; Quan, X.; Ling, Q.; and Song, Y. 2020a. Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation. In *ACL*, 7056–7066. Association for Computational Linguistics.

Li, X.; Bing, L.; Li, P.; Lam, W.; and Yang, Z. 2018. Aspect Term Extraction with History Attention and Selective Transformation. In *IJCAI*, 4194–4200. ijcai.org.

Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; and Li, J. 2020b. A Unified MRC Framework for Named Entity Recognition. In *ACL*, 5849–5859. Association for Computational Linguistics.

Li, X.; and Lam, W. 2017. Deep Multi-Task Learning for Aspect Term Extraction with Memory Interaction. In *EMNLP*, 2886–2892. Association for Computational Linguistics.

Li, X.; Yin, F.; Sun, Z.; Li, X.; Yuan, A.; Chai, D.; Zhou, M.; and Li, J. 2019. Entity-Relation Extraction as Multi-Turn Question Answering. In *ACL (1)*, 1340–1350. Association for Computational Linguistics.

Liu, K.; Xu, H. L.; Liu, Y.; and Zhao, J. 2013. Opinion Target Extraction Using Partially-Supervised Word Alignment Model. In *IJCAI*, 2134–2140. IJCAI/AAAI.

Liu, K.; Xu, L.; and Zhao, J. 2012. Opinion Target Extraction Using Word-Based Translation Model. In *EMNLP-CoNLL*, 1346–1356. ACL.

Liu, K.; Xu, L.; and Zhao, J. 2014. Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking. In *ACL (1)*, 314–324. The Association for Computer Linguistics.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR (Poster)*. OpenReview.net.

Ma, D.; Li, S.; Wu, F.; Xie, X.; and Wang, H. 2019. Exploring Sequence-to-Sequence Learning in Aspect Term Extraction. In *ACL (1)*, 3538–3547. Association for Computational Linguistics.

Mukherjee, A.; and Liu, B. 2012. Aspect Extraction through Semi-Supervised Modeling. In *ACL (1)*, 339–348. The Association for Computer Linguistics.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; Clercq, O. D.; Hoste, V.; Apidianaki, M.; Tannier, X.; Loukachevitch, N. V.; Kotelnikov, E. V.; Bel, N.; Zafra, S. M. J.; and Eryigit, G. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *SemEval@NAACL-HLT*, 19–30. The Association for Computer Linguistics.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; and Androutsopoulos, I. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *SemEval@NAACL-HLT*, 486–495. The Association for Computer Linguistics.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *SemEval@COLING*, 27–35. The Association for Computer Linguistics.

Qiu, G.; Liu, B.; Bu, J.; and Chen, C. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Comput. Linguistics* 37(1): 9–27.

Sun, C.; Huang, L.; and Qiu, X. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In *NAACL-HLT (1)*, 380–385. Association for Computational Linguistics.

Wang, B.; and Wang, H. 2008. Bootstrapping Both Product Features and Opinion Words from Chinese Customer Reviews with Cross-Inducing. In *IJCNLP*, 289–295. The Association for Computer Linguistics.

Wang, W.; and Pan, S. J. 2018. Recursive Neural Structural Correspondence Network for Cross-domain Aspect and Opinion Co-Extraction. In *ACL (1)*, 2171–2181. Association for Computational Linguistics.

Wang, W.; and Pan, S. J. 2019. Transferable Interactive Memory Network for Domain Adaptation in Fine-Grained Opinion Extraction. In *AAAI*, 7192–7199. AAAI Press.

Wang, W.; Pan, S. J.; Dahlmeier, D.; and Xiao, X. 2016. Recursive Neural Conditional Random Fields for Aspect-based Sentiment Analysis. In *EMNLP*, 616–626. The Association for Computational Linguistics.

Wang, W.; Pan, S. J.; Dahlmeier, D.; and Xiao, X. 2017. Coupled Multi-Layer Attentions for Co-Extraction of Aspect and Opinion Terms. In *AAAI*, 3316–3322. AAAI Press.

Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; and Chang, Y. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *ACL*, 1476–1488. Association for Computational Linguistics.

Wu, Z.; Zhao, F.; Dai, X.; Huang, S.; and Chen, J. 2020. Latent Opinions Transfer Network for Target-Oriented Opinion Words Extraction. In *AAAI*, 9298–9305. AAAI Press.

Xu, H.; Liu, B.; Shu, L.; and Yu, P. S. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. In *ACL (2)*, 592–598. Association for Computational Linguistics.

Xu, H.; Liu, B.; Shu, L.; and Yu, P. S. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *NAACL-HLT (1)*, 2324–2335. Association for Computational Linguistics.

Yang, B.; and Cardie, C. 2013. Joint Inference for Fine-grained Opinion Extraction. In *ACL (1)*, 1640–1649. The Association for Computer Linguistics.

Yin, Y.; Wei, F.; Dong, L.; Xu, K.; Zhang, M.; and Zhou, M. 2016. Unsupervised Word and Dependency Path Embeddings for Aspect Term Extraction. In *IJCAI*, 2979–2985. IJCAI/AAAI Press.

Yu, J.; Jiang, J.; and Xia, R. 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.

Zhao, H.; Huang, L.; Zhang, R.; Lu, Q.; and Xue, H. 2020. SpanMlt: A Span-based Multi-Task Learning Framework for Pair-wise Aspect and Opinion Terms Extraction. In *ACL*, 3239–3248. Association for Computational Linguistics.