ASAP: A Chinese Review Dataset Towards Aspect Category Sentiment Analysis and Rating Prediction

Jiahao Bu[1]*, Lei Ren[1]*, Shuang Zheng[1,2]*, Yang Yang[1], Jingang Wang[1]†, Fuzheng Zhang[1], Wei Wu[1]

[1]Meituan, Beijing, China

[2]School of Economics and Management, Dalian University of Technology, Dalian, China

{bujiahao,renlei04,zhengshuang04,yangyang113}@meituan.com

{wangjingang02,zhangfuzheng,wuwei30}@meituan.com

Abstract

Sentiment analysis has attracted increasing attention in e-commerce. The sentiment polarities underlying user reviews are of great value for business intelligence. Aspect category sentiment analysis (ACSA) and review rating prediction (RP) are two essential tasks to detect the fine-to-coarse sentiment polarities. ACSA and RP are highly correlated and usually employed jointly in real-world e-commerce scenarios. While most public datasets are constructed for ACSA and RP separately, which may limit the further exploitations of both tasks. To address the problem and advance related researches, we present a large-scale Chinese restaurant review dataset **ASAP** including $46,730$ genuine reviews from a leading online-to-offline (O2O) e-commerce platform in China. Besides a 5-star scale rating, each review is manually annotated according to its sentiment polarities towards 18 pre-defined aspect categories. We hope the release of the dataset could shed some light on the field of sentiment analysis. Moreover, we propose an intuitive yet effective joint model for ACSA and RP. Experimental results demonstrate that the joint model outperforms state-of-the-art baselines on both tasks.

1 Introduction

With the rapid development of e-commerce, massive user reviews available on e-commerce platforms are becoming valuable resources for both customers and merchants. Aspect-based sentiment analysis(ABSA) on user reviews is a fundamental and challenging task which attracts interests from both academia and industries (Hu and Liu, 2004; Ganu et al., 2009; Jo and Oh, 2011; Kiritchenko et al., 2014). According to whether the aspect terms are explicitly mentioned in texts, ABSA can be further classified into aspect term sentiment analysis (ATSA) and aspect category sentiment analysis (ACSA), we focus on the latter which is more widely used in industries. Specifically, given a review "Although the fish is delicious, the waiter is horrible!", the ACSA task aims to infer the sentiment polarity over aspect category *food* is positive while the opinion over the aspect category *service* is negative.

The user interfaces of e-commerce platforms are more intelligent than ever before with the help of ACSA techniques. For example, Figure 1 presents the detail page of a coffee shop on a popular e-commerce platform in China. The upper aspect-based sentiment text-boxes display the aspect categories (e.g., *food*, *sanitation*) mentioned frequently in user reviews and the aggregated sentiment polarities on these aspect categories (the orange ones represent positive and the blue ones represent negative). Customers can focus on corresponding reviews effectively by clicking the aspect-based sentiment text-boxes they care about (e.g., the orange filled text-box "卫生条件好" (*good sanitation*)). Our user survey based on $7,824$ valid questionnaires demonstrates that 80.08% customers agree that the aspect-based sentiment text-boxes are helpful to their decision-making on restaurant choices. Besides, the merchants can keep track of their cuisines and service qualities with the help of the aspect-based sentiment text-boxes. Most Chinese e-commerce platforms such as Taobao[1], Dianping[2], and Koubei[3] deploy the similar user interfaces to improve user experience.

Users also publish their overall 5-star scale ratings together with reviews. Figure 1 displays a sample of 5-star rating to the coffee shop. In comparison to fine-grained aspect sentiment, the overall review rating is usually a coarse-grained synthesis of the opinions on multiple aspects. Rating pre-

---

*Equal contribution.

†Corresponding author.

[1]https://www.taobao.com/

[2]https://www.dianping.com/

[3]https://www.koubei.com/

diction(RP) (Jin et al., 2016; Li et al., 2018; Wu et al., 2019a) which aims to predict the "seeing stars" of reviews also has wide applications. For example, to promise the aspect-based sentiment text-boxes accurate, unreliable reviews should be removed before ACSA algorithms are performed. Given a piece of user review, we can predict a rating for it based on the overall sentiment polarity underlying the text. We assume the predicted rating of the review should be consistent with its ground-truth rating as long as the review is reliable. If the predicted rating and the user rating of a review disagree with each other explicitly, the reliability of the review is doubtful. Figure 2 demonstrates an example review of low-reliability. In summary, RP can help merchants to detect unreliable reviews.

Therefore, both ACSA and RP are of great importance for business intelligence in e-commerce, and they are highly correlated and complementary. ACSA focuses on predicting its underlying sentiment polarities on different aspect categories, while RP focuses on predicting the user's overall feelings from the review content. We reckon these two tasks are highly correlated and better performance could be achieved by considering them jointly.

As far as we know, current public datasets are constructed for ACSA and RP separately, which limits further joint explorations of ACSA and RP. To address the problem and advance the related researches, this paper presents a large-scale Chinese restaurant review dataset for **A**spect category **S**entiment **A**nalysis and rating **P**rediction, denotes as **ASAP** for short. All the reviews in ASAP are collected from the aforementioned e-commerce platform. There are 46, 730 restaurant reviews attached with 5-star scale ratings. Each review is manually annotated according to its sentiment polarities towards 18 fine-grained aspect categories. To the best of our knowledge, ASAP is the largest Chinese large-scale review dataset towards both ACSA and RP tasks.

We implement several state-of-the-art (SOTA) baselines for ACSA and RP and evaluate their performance on ASAP. To make a fair comparison, we also perform ACSA experiments on a widely used SemEval-2014 restaurant review dataset (Pontiki et al., 2014). Since BERT (Devlin et al., 2018) has achieved great success in several natural language understanding tasks including sentiment analysis (Xu et al., 2019; Sun et al., 2019; Jiang et al., 2019), we propose a joint model that

employs the fine-to-coarse semantic capability of BERT. Our joint model outperforms the competing baselines on both tasks.
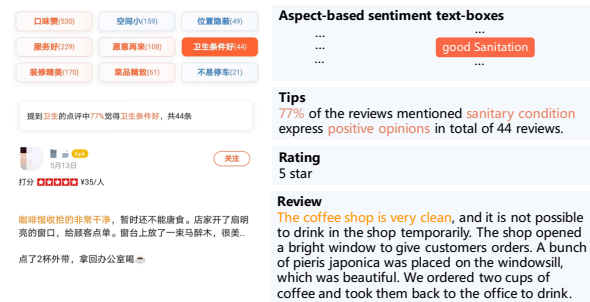

Figure 1: The user interface of a coffee shop on a popular e-commerce App. The top aspect-based sentiment text-boxes display aspect categories and sentiment polarities. The orange text-boxes are positive, while the blue ones are negative. The reviews mentioning the clicked aspect category (e.g., *good sanitation*) with ratings are shown below. The text spans mentioning the aspect categories are also highlighted.
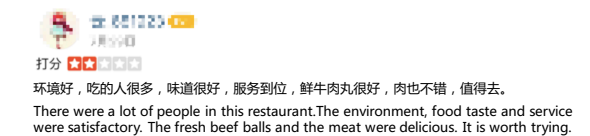

Figure 2: A content-rating disagreement case. The review holds a 2-star rating while all the mentioned aspects are super positive.

Our main contributions can be summarized as follows. (1) We present a large-scale Chinese review dataset towards aspect category sentiment analysis and rating prediction, named as ASAP, including as many as 46, 730 real-world restaurant reviews annotated from 18 pre-defined aspect categories. Our dataset has been released at `https://github.com/Meituan-Dianping/asap`. (2) We explore the performance of widely used models for ACSA and RP on ASAP. (3) We propose a joint learning model for ACSA and RP tasks. Our model achieves the best results both on ASAP and SemEval RESTAURANT datasets.

2   Related Work and Datasets

Aspect   Category   Sentiment   Analysis. ACSA (Zhou et al., 2015; Movahedi et al., 2019; Ruder et al., 2016; Hu et al., 2018) aims to predict sentiment polarities on all aspect categories mentioned in the text. The series of SemEval datasets consisting of user reviews from e-commerce websites have been widely used and

pushed forward related research (Wang et al., 2016; Ma et al., 2017; Xu et al., 2019; Sun et al., 2019; Jiang et al., 2019). The SemEval-2014 task-4 dataset (SE-ABSA14) (Pontiki et al., 2014) is composed of laptop and restaurant reviews. The restaurant subset includes 5 aspect categories (i.e., *Food*, *Service*, *Price*, *Ambience* and *Anecdotes/Miscellaneous*) and 4 polarity labels (i.e., *Positive*, *Negative*, *Conflict* and *Neutral*). The laptop subset is not suitable for ACSA. The SemEval-2015 task-12 dataset (SE-ABSA15) (Pontiki et al., 2015) builds upon SE-ABSA14 and defines its aspect category as a combination of an entity type and an attribute type(e.g., *Food#Style_Options*). The SemEval-2016 task-5 dataset (SE-ABSA16) (Pontiki et al., 2016) extends SE-ABSA15 to new domains and new languages other than English. MAMS (Jiang et al., 2019) tailors SE-ABSA14 to make it more challenging, in which each sentence contains at least two aspects with different sentiment polarities.

Compared with the prosperity of English resources, high-quality Chinese datasets are not rich enough. "ChnSentiCorp" (Tan and Zhang, 2008), "IT168TEST" (Zagibalov and Carroll, 2008), "Weibo"[4], "CTB" (Li et al., 2014) are 4 popular Chinese datasets for general sentiment analysis. However, aspect category information is not annotated in these datasets. Zhao et al. (2014) presents two Chinese ABSA datasets for consumer electronics (mobile phones and cameras). Nevertheless, the two datasets only contain 400 documents ($\sim$ 4000 sentences), in which each sentence only mentions one aspect category at most. BDCI[5] automobile opinion mining and sentiment analysis dataset (Dai et al., 2019) contains 8, 290 user reviews in automobile industry with 10 predefined categories. Peng et al. (2017) summarizes available Chinese ABSA datasets. While most of them are constructed through rule-based or machine learning-based approaches, which inevitably introduce additional noise into the datasets. Our ASAP excels above Chinese datasets both on quantity and quality.

Rating Prediction. Rating prediction (RP) aims to predict the "seeing stars" of reviews, which represent the overall ratings of reviews. In comparison to fine-grained aspect sentiment, the overall review rating is usually a coarse-grained synthesis of the opinions on multiple aspects. Ganu et al. (2009); Li et al. (2011); Chen et al. (2018) form this task as a text classification or regression problem. Considering the importance of opinions on multiple aspects in reviews, recent years have seen numerous work (Jin et al., 2016; Cheng et al., 2018; Li et al., 2018; Wu et al., 2019a) utilizing the information of the aspects to improve the rating prediction performance. This trending also inspires the motivation of ASAP.

Most RP datasets are crawled from real-world review websites and created for RP specifically. Amazon Product Review English dataset (McAuley and Leskovec, 2013) containing product reviews and metadata from Amazon has been widely used for RP (Cheng et al., 2018; McAuley and Leskovec, 2013). Another popular English dataset comes from Yelp Dataset Challenge 2017[6], which includes reviews of local businesses in 12 metropolitan areas across 4 countries. Openrice[7] is a Chinese RP dataset composed of 168, 142 reviews. Both the English and Chinese datasets don't annotate fine-grained aspect category sentiment polarities.

3 Dataset Collection and Analysis

3.1 Data Construction & Curation

We collect reviews from one of the most popular O2O e-commerce platforms in China, which allows users to publish coarse-grained star ratings and writing fine-grained reviews to restaurants (or places of interest) they have visited. In the reviews, users comment on multiple aspects either explicitly or implicitly, including *ambience*,*price*, *food*, *service*, and so on.

First, we retrieve a large volume of user reviews from popular restaurants holding more than 50 user reviews randomly. Then, 4 pre-processing steps are performed to promise the ethics, quality, and reliability of the reviews. (1) User information (e.g., user-ids, usernames, avatars, and post-times) are removed due to privacy considerations. (2) Short reviews with less than 50 Chinese characters, as well as lengthy reviews with more than 1000 Chinese characters are filtered out. (3) If the ratio of non-Chinese characters within a review is over 70%, the review is discarded. (4) To detect the low-

quality reviews (e.g., advertising texts), we build a BERT-based classifier with an accuracy of 97% in a leave-out test-set. The reviews detected as low-quality by the classifier are discarded too.

3.2 Aspect Categories

Since the reviews already hold users' star ratings, this section mainly introduces our annotation details for ACSA. In SE-ABSA14 restaurant dataset (denoted as RESTAURANT for simplicity), there are 5 coarse-grained aspect categories, including *food*, *service*, *price*, *ambience* and *miscellaneous*. After an in-depth analysis of the collected reviews, we find the aspect categories mentioned by users are rather diverse and fine-grained. Take the text "...The restaurant holds a high-end decoration but is quite noisy since a wedding ceremony was being held in the main hall... (...环境看起来很高大上的样子，但是因为主厅在举办婚礼非常混乱，感觉特别吵...)" in Table 3 for example, the reviewer actually expresses opposite sentiment polarities on two fine-grained aspect categories related to *ambience*. The restaurant's decoration is very high-end (*Positive*), while it's very noisy due to an ongoing ceremony (*Negative*). Therefore, we summarize the frequently mentioned aspects and refine the 5 coarse-grained categories into 18 fine-grained categories. We replace *miscellaneous* with *location* since we find users usually review the restaurants' location (e.g., whether the restaurant is easy to reach by public transportation.). We denote the aspect category as the form of "*Coarse-grained Category#Fine-grained Categoty*", such as "*Food#Taste*" and "*Ambience#Decoration*". The full list of aspect categories and definitions are listed in Table 1.

3.3 Annotation Guidelines & Process

Bearing in mind the pre-defined 18 aspects, assessors are asked to annotate sentiment polarities towards the mentioned aspect categories of each review. Given a review, when an aspect category is mentioned within the review either explicitly and implicitly, the sentiment polarity over the aspect category is labeled as 1 (*Positive*), 0 (*Neutral*) or −1 (*Negative*) as shown in Table 3.

We hire 20 vendor assessors, 2 project managers, and 1 expert reviewer to perform annotations. Each assessor needs to attend a training to ensure their intact understanding of the annotation guidelines. Three rounds of annotation are conducted sequentially. First, we randomly split the whole dataset

into 10 groups, and every group is assigned to 2 assessors to annotate independently. Second, each group is split into 2 subsets according to the annotation results, denoted as **Sub-Agree** and **Sub-Disagree**. **Sub-Agree** comprises the data examples with agreement annotation, and **Sub-Disagree** comprises the data examples with disagreement annotation. **Sub-Agree** will be reviewed by assessors from other groups. The controversial examples during the review are considered as difficult cases. **Sub-Disagree** will be reviewed by the 2 project managers independently and then discuss to reach an agreement annotation. The examples that could not be addressed after discussions are also considered as difficult cases. Third, for each group, the difficult examples from two subsets are delivered to the expert reviewer to make a final decision. More details of difficult cases and annotation guidelines during annotation are demonstrated in Table 2.

Finally, ASAP corpus consists of 46,730 pieces of real-world user reviews, and we split it into a training set (36,850), a validation set (4,940) and a test set (4,940) randomly. Table 3 presents an example review of ASAP and corresponding annotations on the 18 aspect categories.

3.4 Dataset Analysis

Figure 3 presents the distribution of 18 aspect categories in ASAP. Because ASAP concentrates on the domain of restaurant, 94.7% reviews mention *Food#Taste* as expected. Users also pay great attention to aspect categories such as *Service#Hospitality*, *Price#Level* and *Ambience#Decoration*. The distribution proves the advantages of ASAP, as users' fine-grained preferences could reflect the pros and cons of restaurants more precisely.

The statistics of ASAP are presented in Table 4. We also include a tailored SE-ABSA14 RESTAURANT dataset for reference. Please note that we remove the reviews holding aspect categories with sentiment polarity of "conflict" from the original RESTAURANT dataset.

Compared with RESTAURANT, ASAP excels in the quantities of training instances, which supports the exploration of recent data-intensive deep neural models. ASAP is a review-level dataset, while RESTAURANT is a sentence-level dataset. The average length of reviews in ASAP is much longer, thus the reviews tend to contain richer aspect information. In ASAP, the reviews contain

Table 1: The full list of 18 aspect categories and definitions.

| Aspect category	Definition	Aspect category	Definition
Food#Taste (口味)	Food taste	*Location#Easy_to_find* (是否容易寻找)	Whether the restaurant is easy to find
Food#Appearance (外观)	Food appearance	*Service#Queue* (排队时间)	Whether the queue time is acceptable
Food#Portion (分量)	Food portion	*Service#Hospitality* (服务人员态度)	Waiters/waitresses' attitude/hospitality
Food#Recommend (推荐程度)	Whether the food is worth being recommended	*Service#Parking* (停车方便)	Parking convenience
Price#Level (价格水平)	Price level	*Service#Timely* (点菜/上菜速度)	Order/Serving time
Price#Cost_effective (性价比)	Whether the restaurant is cost-effective	*Ambience#Decoration* (装修)	Decoration level
Price#Discount (折扣力度)	Discount strength	*Ambience#Noise* (嘈杂情况)	Whether the restaurant is noisy
Location#Downtown (位于商圈附近)	Whether the restaurant is located near downtown	*Ambience#Space* (就餐空间)	Dining Space and Seat Size
Location#Transportation (交通方便)	Convenient public transportation to the restaurant	*Ambience#Sanitary* (卫生情况)	Sanitary condition
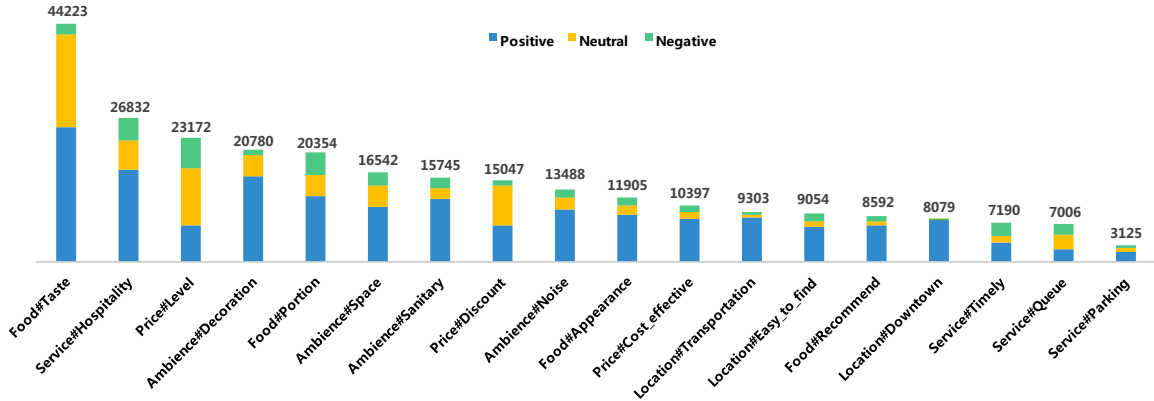
Figure 3: The distribution of 18 fine-grained aspect categories in ASAP.

5.8 aspect categories in average, which is 4.7 times of RESTAURANT. Both review-level ACSA and RP are more challenging than their sentence-level counterparts. Take the review in Table 3 for example, the review contains several sentiment polarities towards multiple aspect categories. In addition to aspect category sentiment annotations, ASAP also includes overall user ratings for reviews. With the help of ASAP, ACSA and RP can be further optimized either separately or jointly.

4 Methodology

4.1 Problem Formulation

We use D to denote the collection of user review corpus in the training data. Given a review R which consists of a series of words: $\{w_1, w_2, ..., w_Z\}$, ACSA aims to predict the sentiment polarity $y_i \in \{Positive, Neutral, Negative\}$ of review R with respect to the mentioned aspect category a_i, $i \in \{1, 2, ..., N\}$. Z denotes the length of review R. N is the number of pre-defined aspect categories (i.e., 18 in this paper). Suppose there are K mentioned aspect categories in R. We define a mask vector $[p_1, p_2, ..., p_N]$ to indicate the occurrence of aspect categories. When the aspect category a_i is mentioned in R, $p_i = 1$, otherwise $p_i = 0$. So we have $\sum_{i=1}^{N} p_i = K$. In terms of RP, it aims to predict the 5-star rating score of g, which represents the overall rating of the given review R.

4.2 Joint Model

Given a user review, ACSA focuses on predicting its underlying sentiment polarities on different aspect categories, while RP focuses on predicting the user's overall feelings from the review content. We reckon these two tasks are highly correlated and better performance could be achieved by considering them jointly.

The advent of BERT has established the success

Table 2: Difficult cases and annotation guidelines.

Category	Example	Example (Translation)	Guideline	Annotation
Change of sentiment over time	我之前挺喜欢这家餐厅的饭菜，不过今天的饭菜可不怎么样.	I used to like the food of this restaurant, but the taste is not as expected today.	When there existed a sentiment drifting over time in the review, the most recent sentiment polarity is adopted.	(*Food#Taste*, −1)
Implicit sentiment polarity	比五星级酒店的餐厅差远了，而且在五星级酒店中餐厅里两个人吃一顿也就500左右就够了	The restaurant was far worse than the dinning hall of any five-star hotel, considering that the meal for two people only cost 500 CNY in a five-star hotel.	Some reviewers express their polarities in an implicit manner instead of expressing their feelings directly. The implicit sentiment polarity is adopted to complete the annotation.	(*Price#Level*, −1)
Conflict opinions	这道菜有点咸，但是味道很赞。	This dish was a bit salty, but it tasted great.	When there existed multiple sentiment polarities toward the same aspect-category, the dominant sentiment is chosen.	(*Food#Taste*, 1)
Mild sentiment	饭菜还可以，不过也算不上特别好吃。	The food was okay, but nothing great.	The "neutral" label applies to mildly positive or mildly negative sentiment	(*Food#Taste*, 0)
Irrelevant restaurant	上次去的一家店很难吃，今天来了这家新的，感觉很好吃。	The food of the shop which I went to last time was very bad. Today I came to this new one. I felt very good.	The review mentions restaurant that the user has visited in the past. We only focus on the restaurant being reviewed	(*Food#Taste*, 1)

of the "pre-training and then fine-tuning" paradigm for NLP tasks. BERT-based models have achieved impressive results in ACSA (Xu et al., 2019; Sun et al., 2019; Jiang et al., 2019). Review rating prediction can be deemed as a single-sentence classification (regression) task, which could also be addressed with BERT. Therefore, we propose a joint learning model to address ACSA and RP in a multitask learning manner. Our joint model employs the fine-to-coarse semantic representation capability of the BERT encoder. Figure 4 illustrates the framework of our joint model.

ACSA As shown in Figure 4, the token embeddings of the input review are generated through a shared BERT encoder. Briefly, let $H \in \mathbb{R}^{d*Z}$ be the matrix consisting of token embedding vectors $\{h_1, ..., h_Z\}$ that BERT produces, where d is the size of hidden layers and Z is the length of the given review. Since different aspect category information is dispersed across the content of R, we add an attention-pooling layer (Wang et al., 2016) to aggregate the related token embeddings dynamically for every aspect category. The attention-pooling layer helps the model focus on the tokens most related to the target aspect categories.

$$M_i^a = \tanh(W_i^a * H) \qquad (1)$$

$$\alpha_i = softmax(\omega_i^T * M_i^a) \qquad (2)$$

$$r_i = \tanh(W_i^p * H * \alpha_i^T) \qquad (3)$$

Where $W_i^a \in \mathbb{R}^{d*d}$, $M_i^a \in \mathbb{R}^{d*Z}$, $\omega_i \in \mathbb{R}^d$, $\alpha_i \in \mathbb{R}^Z$, $W_i^p \in \mathbb{R}^{d*d}$, and $r_i \in \mathbb{R}^d$. α_i is a vector consisting of attention weights of all tokens which can selectively attend the regions of the aspect category related tokens, and r_i is the attentive representation of review with respect to the i_{th} aspect category a_i, $i \in \{1, 2, ..., N\}$. Then we have

$$\hat{y}_i = softmax(W_i^q * r_i + b_i^q) \qquad (4)$$

Where $W_i^q \in \mathbb{R}^{C*d}$ and $b_i^q \in \mathbb{R}^C$ are trainable parameters of the softmax layer. C is the number of labels (i.e, 3 in our task). Hence, the ACSA loss for a given review R is defined as follows,

$$loss_{ACSA} = \frac{1}{K} \sum_{i=1}^{N} p_i \sum_{C} y_i * \log \hat{y}_i \qquad (5)$$

If the aspect category a_i is not mentioned in S, y_i is set as a random value. The p_i serves as a gate function, which filters out the random y_i and ensures only the mentioned aspect categories can participate in the calculation of the loss function.

Table 3: A review example in ASAP, with overall star rating and aspect category sentiment polarity annotations.

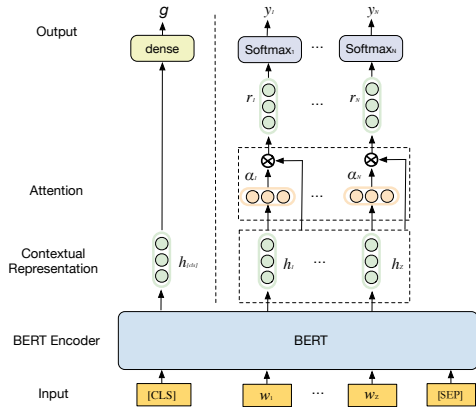| Review	Rating	Aspect Category	Label	Aspect Category	Label
With convenient traffic, the restaurant holds a high-end decoration, but quite noisy because a wedding ceremony was being held in the main hall. Impressed by its delicate decoration and grand appearance though, we had to wait for a while at the weekend time. However, considering its high price level, the taste is unexpected. We ordered the Kung Pao Prawn, the taste was acceptable and the serving size is enough, but the shrimp is not fresh. In terms of service, you could not expect too much due to the massive customers there. By the way, the free-served fruit cup was nice. Generally speaking, it was a typical wedding banquet restaurant rather than a comfortable place to date with friends. 交通还挺方便的，环境看起来很高大上的样子，但是因为主厅在举办婚礼非常混乱，特别吵感觉，但是装修的还不错，感觉很精致的装修，门面很气派，周末去的时候还需要等位。味道的话我觉得还可以但是跟价格比起来就很一般了，性价比挺低的，为了去吃宫保虾球的，但是我觉得也就那样吧虾不是特别新鲜，不过虾球很大，味道还行。服务的话由于人很多所以也顾不过来上菜的速度不快，但是有送水果杯还挺好吃的。总之就是典型的婚宴餐厅不是适合普通朋友吃饭的地方了。	3-Star	Location#Transportation (交通方便)	1	Price#Discount (折扣力度)	-
		Location#Downtown (位于商圈附近)	-	Ambience#Decoration (装修)	1
		Location#Easy_to_find (是否容易寻找)	-	Ambience#Noise (嘈杂情况)	−1
		Service#Queue (排队时间)	-	Ambience#Space (就餐空间)	1
		Service#Hospitality (服务人员态度)	-	Ambience#Sanitary (卫生情况)	-
		Service#Parking (停车方便)	-	Food#Portion (分量)	1
		Service#Timely (点菜/上菜速度)	−1	Food#Taste (口味)	1
		Price#Level (价格水平)	0	Food#Appearance (外观)	-
		Price#Cost_effective (性价比)	−1	Food#Recommend (推荐程度)	-

Figure 4: The framework of the proposed joint learning model. The right part of the dotted vertical line is used to predict multiple aspect category sentiment polarities, while the left part is used to predict the review rating.

Rating Prediction Since the objective of RP is to predict the review rating based on the review content, we adopt the [CLS] embedding $h_{[cls]} \in \mathbb{R}^d$

BERT produces as the representation of the input review, where d is the size of hidden layers in the BERT encoder.

$$\hat{g} = \beta^T * \tanh(W^r * h_{[cls]} + b^r) \qquad (6)$$

Hence the RP loss for a given review R is defined as follows,

$$loss_{RP} = |g - \hat{g}| \qquad (7)$$

Where $W^r \in \mathbb{R}^{d*d}, b^r \in \mathbb{R}^d$, $\beta \in \mathbb{R}^d$ are trainable parameters.

The final loss of our joint model becomes as follows.

$$loss = loss_{ACSA} + loss_{RP} \qquad (8)$$

5 Experiments

We perform an extensive set of experiments to evaluate the performance of our joint model on ASAP

2075

Table 4: The statistics and label/rating distribution of ASAP and RESTAURANT. The review length are counted by Chinese characters and English words respectively. The sentences are segmented with periods in ASAP, while RESTAURANT is a sentence-level dataset.

Dataset	Split	Reviews	Average sentences per review	Average aspects per review	Average length	Positive	Negative	Neutral	1-star	2-star	3-star	4-star	5-star
ASAP	Train	36,850	8.6	5.8	319.7	133,721	27,425	52,225	1,219	1,258	5,241	13,362	15,770
	Dev	4,940	8.7	5.9	319.9	18,176	3,733	7,192	151	166	784	1,734	2,105
	Test	4,940	8.3	5.7	317.1	17,523	3,813	7,026	165	173	717	1,867	2,018
RESTAURANT	Train	2,855	1	1.2	15.2	2150	822	498	-	-	-	-	-
	Test	749	1	1.3	15.6	645	215	94	-	-	-	-	-

and RESTAURANT (Pontiki et al., 2014). Ablation studies are also conducted to probe the interactive influence between ACSA and RP.

5.1 ACSA

Baseline Models We implement several ACSA baselines for comparison. According to the different structures of their encoders, these models are classified into Non-BERT based models or BERT-based models. Non-BERT based models include TextCNN (Kim, 2014), BiLSTM+Attn (Zhou et al., 2016), ATAE-LSTM (Wang et al., 2016) and CapsNet (Sabour et al., 2017). BERT-based models include vanilla BERT (Devlin et al., 2018), QA-BERT (Sun et al., 2019) and CapsNet-BERT (Jiang et al., 2019).

Implementation Details of Experimental Models In terms of non-BERT-based models, we initialize their inputs with pre-trained embeddings. For Chinese ASAP, we utilize Jieba[8] to segment Chinese texts and adopt Tencent Chinese word embeddings (Song et al., 2018) composed of $8,000,000$ words. For English RESTAURANT, we adopt 300-dimensional word embeddings pre-trained by Glove (Pennington et al., 2014).

In terms of BERT-based models, we adopt the 12-layer Google BERT Base[9] to encode the inputs.

The batch sizes are set as 32 and 16 for non-BERT-based models and BERT-based models respectively. Adam optimizer (Kingma and Ba, 2014) is employed with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The maximum sequence length is set as 512. The number of epochs is set as 3. The learning rates are set as 0.001 and 0.00005 for non-BERT-based models and BERT-based models respectively. All the models are trained on a single NVIDIA Tesla 32G V100 Volta GPU.

Evaluation Metrics Following the settings of RESTAURANT, we adopt Macro-F1 and Accuracy

---

[8] https://github.com/fxsjy/jieba
[9] https://github.com/google-research/bert

(Acc) as evaluation metrics.

Experimental Results & Analysis We report the performance of aforementioned models on ASAP and RESTAURANT in Table 5. Generally, BERT-based models outperform Non-BERT based models on both datasets. The two variants of our joint model perform better than vanilla-BERT, QA-BERT, and CapsNet-BERT, which proves the advantages of our joint learning model. Given a user review, vanilla-BERT, QA-BERT, and CapsNet-BERT treat the pre-defined aspect categories independently, while our joint model combines them together with a multi-task learning framework. On one hand, the encoder-sharing setting enables knowledge transferring among different aspect categories. On the other hand, our joint model is more efficient than other competitors, especially when the number of aspect categories is large. The ablation of RP (i.e., joint model(w/o RP)) still outperforms all other baselines. The introduction of RP to ACSA brings marginal improvement. This is reasonable considering that the essential objective of RP is to estimate the overall sentiment polarity instead of fine-grained sentiment polarities.

We visualize the attention weights produced by our joint model on the example of Table 3 in Figure 5. Since different aspect category information is dispersed across the review of R, we add an attention-pooling layer (Wang et al., 2016) to aggregate the related token embeddings dynamically for every aspect category. The attention-pooling layer helps the model focus on the tokens most related to the target aspect categories. Figure 5 visualizes attention weights of 3 given aspect categories. The intensity of the color represents the magnitude of attention weight, which means the relatedness of tokens to the given aspect category. It's obvious that our joint model focus on the tokens most related to the aspect categories across the review of R.

Table 5: The experimental results of ACSA models on ASAP and RESTAURANT. Best scores are boldfaced.

Category	Model	ASAP		RESTAURANT	
		Macro-F1	Acc.	Macro-F1	Acc.
Non-BERT-based models	TextCNN (Kim, 2014)	60.41%	71.10%	70.56%	82.29%
	BiLSTM+Attn (Zhou et al., 2016)	70.53%	77.78%	70.85%	81.97%
	ATAE_LSTM (Wang et al., 2016)	76.60%	81.94%	70.15%	82.12%
	CapsNet (Sabour et al., 2017)	75.54%	81.66%	71.84%	82.63%
BERT-based models	Vanilla-BERT (Devlin et al., 2018)	79.18%	84.09%	79.22%	87.63%
	QA-BERT (Sun et al., 2019)	79.44%	83.92%	80.89%	88.89%
	CapsNet-BERT (Jiang et al., 2019)	78.92%	83.74%	80.94%	89.00%
	Joint Model (w/o RP)	80.75%	85.15%	**82.01%**	**89.62%**
	Joint Model	**80.78%**	**85.19%**	-	-

Review:

交通还挺方便的，环境看起来很高大上的样子，但是因为主厅在举办婚礼非常混乱，特别吵感觉，但是装修的还不错，感觉很精致的装修，门面很气派，周末去的时候还需要等位。味道的话我觉得还可以但是跟价格比起来就很一般了，性价比挺低的，为了去吃宫保虾球的，但是我觉得也就那样吧虾不是特别新鲜，不过虾球很大，味道还行。服务的话由于人很多所以也顾不过来上菜的速度不快，但是有送水果杯还挺好吃的。总之就是典型的婚宴餐厅不是适合普通朋友吃饭的地方了。

With convenient traffic, the restaurant holds a high-end decoration, but quite noisy because a wedding ceremony was being held in the main hall. Impressed by its delicate decoration and grand appearance though, we had to wait for a while at the weekend time. However, considering its high price level, the taste is unexpected. We ordered the Kung Pao Prawn, the taste was acceptable and the serving size is enough, but the shrimp is not fresh. In terms of service, you could not expect too much due to the massive customers there. By the way, the free-served fruit cup was nice. Generally speaking, it was a typical wedding banquet restaurant rather than a comfortable place to date with friends.

Figure 5: Attention visualization example. We only show attention weights of 3 aspect categories for beauty. The red text span "..With convenient traffic..(..交通还挺方便的..)" is related to *Location#Transportation*. The blue text span "..the restaurant holds a high-end decoration..Impressed by its delicate decoration and grand appearance though..(..环境看起来高大上的样子..装修还不错，很精致的装修..)" is related to *Ambience#Decoration*. The green text span "..but quite noisy..(..特别吵感觉..)" is related to *Ambience#Noise*. The intensity of the color represents the magnitude of attention weight.

5.2 Rating Prediction

We compare several RP models on ASAP, including TextCNN (Kim, 2014), BiLSTM+Attn (Zhou et al., 2016) and ARP (Wu et al., 2019b). The data pre-processing and implementation details are identical with ACSA experiments.

Evaluation Metrics. We adopt Mean Absolute Error (MAE) and Accuracy (by mapping the predicted rating score to the nearest category) as evaluation metrics.

Experimental Results & Analysis The experimental results of comparative RP models are illustrated in Table 6.

Table 6: Experimental results of RP models on ASAP. Best scores are boldfaced.

Model	MAE	Acc.
TextCNN (Kim, 2014)	.5814	52.99%
BiLSTM+Attn (Zhou et al., 2016)	.5737	54.38%
ARP (Wu et al., 2019b)	.5620	54.76%
Joint Model (w/o ACSA)	.4421	60.08%
Joint Model	**.4266**	**61.26%**

Our joint model which combines ACSA and RP outperforms other models considerably. On one hand, the performance improvement is expected since our joint model is built upon BERT. On the other hand, the ablation of ACSA (i.e., joint model(w/o ACSA)) brings performance degradation of RP on both metrics. We can conclude that the fine-grained aspect category sentiment prediction of the review indeed helps the model predict its overall rating more accurately.

This section conducts preliminary experiments to evaluate classical ACSA and RP models on our proposed ASAP dataset. We believe there still exists much room for improvements to both tasks, and we will leave them for future work.

6 Conclusion

This paper presents **ASAP**, a large-scale Chinese restaurant review dataset towards aspect category sentiment analysis (ACSA) and rating prediction (RP). ASAP consists of 46, 730 restaurant user reviews with star ratings from a leading e-commerce platform in China. Each review is manually annotated according to its sentiment polarities on 18 fine-grained aspect categories. Besides evaluations of ACSA and RP models on ASAP separately, we also propose a joint model to address ACSA and RP synthetically, which outperforms other state-of-the-art baselines considerably. we hope the release of ASAP could push forward related researches and applications.

References

Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1583–1592.

Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. 2018. Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proceedings of the 2018 world wide web conference*, pages 639–648.

Zehui Dai, Wei Dai, Zhenhua Liu, Fengyun Rao, Huajie Chen, Guangpeng Zhang, Yadong Ding, and Jiyang Liu. 2019. Multi-task multi-head attention memory network for fine-grained sentiment analysis. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 609–620. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6. Citeseer.

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2018. CAN: constrained attention networks for multi-aspect sentiment analysis. *CoRR*, abs/1812.10735.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based 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)*, pages 6281–6286.

Zhipeng Jin, Qiudan Li, Daniel D Zeng, YongCheng Zhan, Ruoran Liu, Lei Wang, and Hongyuan Ma. 2016. Jointly modeling review content and aspect ratings for review rating prediction. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 893–896.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882.*

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442.

Changliang Li, Bo Xu, Gaowei Wu, Saike He, Guanhua Tian, and Hongwei Hao. 2014. Recursive deep learning for sentiment analysis over social data. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 180–185. IEEE.

Fangtao Li, Nathan Nan Liu, Hongwei Jin, Kai Zhao, Qiang Yang, and Xiaoyan Zhu. 2011. Incorporating reviewer and product information for review rating prediction. In *Twenty-second international joint conference on artificial intelligence*.

Junjie Li, Haitong Yang, and Chengqing Zong. 2018. Document-level multi-aspect sentiment classification by jointly modeling users, aspects, and overall ratings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 925–936, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *International Joint Conference on Artificial Intelligence*, pages 4068–4074.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Sajad Movahedi, Erfan Ghadery, Heshaam Faili, and Azadeh Shakery. 2019. Aspect category detection via topic-attention network.

Haiyun Peng, Erik Cambria, and Amir Hussain. 2017. A review of sentiment analysis research in chinese language. *Cognitive Computation*, 9(4):423–435.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment

analys. In *International Workshop on Semantic Evaluation (SemEval 2015)*.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation (SemEval 2015)*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. A hierarchical model of reviews for aspect-based sentiment analysis.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*.

Songbo Tan and Jin Zhang. 2008. An empirical study of sentiment analysis for chinese documents. *Expert Systems with applications*, 34(4):2622–2629.

Yequan Wang, Minlie Huang, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Empirical Methods in Natural Language Processing*, pages 606–615.

Chuhan Wu, Fangzhao Wu, Junxin Liu, Yongfeng Huang, and Xing Xie. 2019a. Arp: Aspect-aware neural review rating prediction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2169–2172.

Chuhan Wu, Fangzhao Wu, Junxin Liu, Yongfeng Huang, and Xing Xie. 2019b. Arp: Aspect-aware neural review rating prediction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2169–2172.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Taras Zagibalov and John A Carroll. 2008. Unsupervised classification of sentiment and objectivity in chinese text. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Yanyan Zhao, Bing Qin, and Ting Liu. 2014. Creating a fine-grained corpus for chinese sentiment analysis. *IEEE Intelligent Systems*, 30(1):36–43.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015. Representation learning for aspect category detection in online reviews. In *Twenty-ninth Aaai Conference on Artificial Intelligence*.