Information Processing and Management 58 (2021) 102729

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Aspect-Based Pair-Wise Opinion Generation in Chinese automotive reviews: Design of the task, dataset and model

Yijiang Liu, Fei Li, Donghong Ji *

Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China

ARTICLE INFO

Keywords:
Fine-grained opinion analysis
Aspect-based generation
Pair-wise generation
Joint model

ABSTRACT

The extraction of opinion target–word pairs from user reviews has received much attention recently, since it can provide essential information for fine-grained opinion analysis. However, according to our statistics on a large-scale dataset of Chinese reviews, about 60% reviews do not explicitly show opinion targets or words. To investigate this problem, this paper introduces a new task under fine-grained opinion analysis, named **A**spect-**B**ased **P**air-wise **O**pinion **G**eneration (ABPOG), which aims to generate opinion target–word pairs based on reviews and aspects. To perform this task, we develop a sequence-to-sequence model for opinion target–word pair generation by extending the pointer–generator network with two approaches: (1) an aspect-aware encoder that receives an additional aspect embedding as input to extract aspect-specific features, (2) two hierarchical decoders including a token-level GRU and a global GRU to generate opinion targets and words jointly. To empirically evaluate our task and model, we develop a multi-aspect dataset for ABPOG based on Chinese automotive reviews. Extensive experiments on our dataset show that our model outperforms several strong baselines adapted from the state-of-the-art aspect-based summarization method.

1. Introduction

User-generated reviews express their opinions on used products in multiple aspects, which are valuable to potential consumers and product manufacturers. Since reviews are usually not concise, it takes much time to find out key points from them.

Opinion analysis aims to obtain the opinions of certain specific aspects or targets in sentence- or document-level, including multiple subtasks like aspect detection (Bagheri, 2019; Cavalcanti & Prudêncio, 2017; He, Lee, Ng, & Dahlmeier, 2017; Kushwaha & Chaudhary, 2017; Li, Zhou, & Li, 2015; Luo, Huang, & Zhu, 2019), aspect-level sentiment classification (Fan, Feng, & Zhao, 2018; Li, Bing, Lam, & Shi, 2018a; Meskele & Frasincar, 2020; Tang, Qin, & Liu, 2016; Varghese & Jayasree, 2013), and aspect-based summarization (Frermann & Klementiev, 2019; Krishna & Srinivasan, 2018; Kunneman, Wubben, van den Bosch, & Krahmer, 2018; Mukherjee et al., 2020). As more fine-grained opinion information is needed, fine-grained opinion analysis has been proposed which consists of multiple tasks like subjective expressions detection, polarity/intensity classification, opinion target identification, and opinion holder identification (Diaz, Zhang, & Ng, 2020; Liu, Joty, & Meng, 2015; Wiebe, Wilson, & Cardie, 2005).

Recently, some work aims to extract the opinion target–word pairs from reviews (Dai & Song, 2019; Luo, Li, Liu, Wang & Unger, 2019; Wan et al., 2020; Xu, Liu, Shu, & Yu, 2018) in order to collect fine-grained information. Opinion targets (OT), also called aspect terms, are the words in the sentence that represent the entity or subject evaluated by users. Opinion words (OW), sometimes called opinion terms, are the words used to express the user's attitude to an opinion target. As shown in the first example of Table 1,

* Corresponding author.
E-mail addresses: cslyj@whu.edu.cn (Y. Liu), dhji@whu.edu.cn (F. Li), foxlf823@gmail.com (D. Ji).

https://doi.org/10.1016/j.ipm.2021.102729
Received 1 March 2021; Received in revised form 12 August 2021; Accepted 12 August 2021
Available online 14 September 2021

Table 1
Examples of opinion target–word pairs. The upper part contains an automobile review and its translation, and the lower part shows there groups of aspects (e.g., comfort), opinion targets (OT) (e.g., 减震) and opinion words (OW) (e.g., 不错).

Review: 车子的减震不错,开起来很舒适。本人身高180cm，坐第三排也没问题。车子开起来有很强的推背感　高速上超车毫不费劲 Translation: The car has good shock absorption and is very comfortable to drive. I am 180 cm tall, and I have no problem sitting in the third row. I can get a gratifying jump forward when driving the car, so it does not take much effort to overtake at high speed.
comfort: 减震 (shock absorption) - 不错 (good) *space*: 第三排空间(space of the third row) - 够用 (enough) *power*: 加速 (acceleration) - 强劲 (strong)

"减震" (shock absorption) is an opinion target of the aspect category "*comfort*", and the corresponding opinion word is "不错" (good).

According to our statistics on a large-scale dataset, OTs and OWs are not explicitly present in about 60% of Chinese reviews. Instead, they are implicitly entailed in the text of reviews. As shown in the second and third examples of Table 1, "第三排空间" (space in the third row) and "加速" (acceleration) are OTs, "够用" (enough) and "强劲" (strong) are OWs, corresponding to the aspect categories "*space*" and "*power*", respectively. All of the above OTs and OWs do not present in the review exactly. Instead, we can only infer them from the context in the review.

The current research mainly focuses on the OTs and OWs that exist in the reviews (Wan et al., 2020), and pays less attention to those that do not. To fill the gap in the existing literature, we propose a new task, named **A**spect-**B**ased **P**air-wise **O**pinion **G**eneration (ABPOG), which aims to generate a corresponding ⟨OT, OW⟩ pair given a review and an aspect category. We build our model based on the pointer–generator network (See, Liu, & Manning, 2017) that generate outputs by extracting original words from inputs and generating new words from vocabularies simultaneously, inspired by the success of the sequence-to-sequence framework in keyphrase generation and summarization (Liu & Lapata, 2019; Mehta & Majumder, 2018; Meng et al., 2017; Rush, Chopra, & Weston, 2015; Yuan et al., 2020).

Moreover, to adapt the pointer–generator network to our task, we propose two approaches to enhance it. First, since our task requires the model to generate different opinion target–word pairs from the same review for different aspect categories, aspect-specific features should be taken into consideration. To this end, we propose an **A**spect-aware **G**ated **R**ecurrent **U**nit (AGRU), which transforms the token feature according to the given aspect category in every encoding step. Second, to generate compatible opinion target–word pairs, we treat the opinion target and word generation as a joint process by decoding them into one target–word pair sequence via two correlated hierarchical decoders.

To evaluate our task and model, we construct a large-scale multi-aspect dataset with the help of the Stanford Constituency Parser for ABPOG, in which the data come from a Chinese automobile forum. In the dataset, each review expresses multiple aspects category and each category corresponds to one target–word pair. We evaluate our model based on the dataset from multiple groups of experiments: (1) Compared with several strong baselines adapted from the state-of-the-art aspect-based summarization method (Frermann & Klementiev, 2019), our model outperforms them by about 2%–3% in macro recall. (2) Given that the same attitude of a user can be expressed in various synonyms, the evaluation metrics based on exact string matching cannot reflect performances in a comprehensive view. We propose an evaluation metric based on semantic similarity to support semantic-level fuzzy string matching. Results show that our model also outperforms the baselines by about 1%–2% in macro recall. (3) We also conduct manual analysis for our model via case studies. Results demonstrate that our model can generate reasonable opinion target–word pairs and explainable attention weights for input words with regards to different aspects.

Objective. Our research objectives mainly include the following two aspects. First, we aimed to research aspect-based opinion target–word pairs generation in Chinese auto reviews. With further research in the field of fine-grained opinion analysis and the application of practical needs, the extraction of opinion target–word pairs has gained more attention in e-commerce reviews, which was solved by extraction-based methods. However, some OTs and OWs which do not appear in the review text but are contained in the expression, cannot be handled by the extraction-based method. Our work aims to explore the joint generation of opinion target–word pairs. Moreover, from a practical point of view, we use aspects as extra input to eliminate unnecessary pairs since we only need to obtain the opinion pairs of certain aspects. To our best knowledge, our work is the first to focus on the co-generation of absent opinion target–word pairs in Chinese auto reviews.

Second, we aim to provide data and methods for those who use computational linguistics to mine opinions. With the continuous expansion of corpus, fine-grained opinion analysis based on deep learning methods has made great progress. sequence-to-sequence framework has been proved effective in text generation tasks, but it is rarely used in co-generation under the limitation of aspect. Therefore, we construct a relevant dataset and propose a benchmark model based on the sequence-to-sequence framework.

In summary, we make the following contributions:

- We introduce a new task **ABPOG** and construct a large-scale multi-aspect dataset, which can be used as a benchmark for future research related to aspect-based sentiment analysis, opinion target–word extraction, or generation.
- We build a sequence-to-sequence model based on the pointer–generator network for ABPOG, and propose two effective approaches to enhance our model, namely AGRU-based encoding and jointly hierarchical decoding.

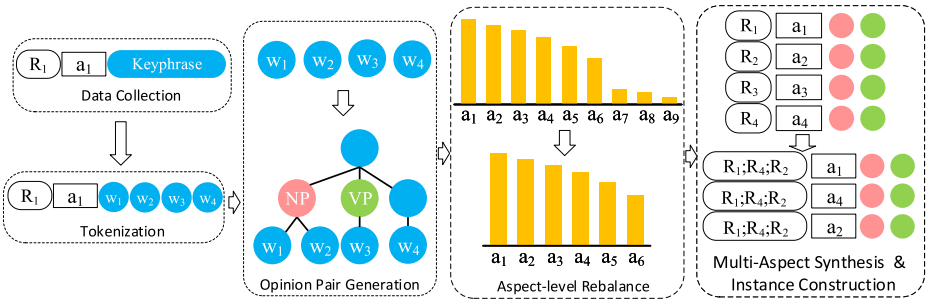

Fig. 1. Steps of dataset construction. R denotes reviews published by users, and a is the aspect category that corresponded to R.

- We conduct extensive experiments and analyses on the constructed dataset. Results show that our model can significantly outperform the baselines, generate reasonable results and meanwhile hold certain interpretability.
- All the codes and datasets will be open-sourced under Apache License 2.0 at https://github.com/liuyijiang1994/ABPOG.

2. Related work

Opinion targets and opinion words are two components in the task of fine-grained opinion analysis. Many studies have investigated this important task (Akhtar, Garg, & Ekbal, 2020; Li, Bing, Li, Lam, & Yang, 2018b; Liao et al., 2019; Ma, Li, Wu, Xie, & Wang, 2019; Tao & Zhou, 2020; Yin, Wu, & Chang, 2020). Some studies deal with the co-extraction of opinion targets and words (Wang & Pan, 2019; Wang, Pan, Dahlmeier, & Xiao, 2016, 2017; Yu, Jiang, & Xia, 2019). For example, Fan, Wu, Dai, Huang, and Chen (2019) treated the opinion target and opinion word as a pair, and proposed a task of target-oriented opinion word extraction, which aims to extract the opinion word for a given target. Chen, Liu, Wang, Zhang, and Chi (2020b) extracted opinion targets and words simultaneously using a synchronous double-channel recurrent network. Different from the above work, Zhao, Huang, Zhang, Lu, and Xue (2020) reconsidered opinion target–word pair extraction from a new perspective of joint term and relation extraction. Peng, Xu, Bing, Huang, Lu, and Si (2020) further proposed a two-stage framework by adding sentiments into opinion word–target pairs.

Besides the research about model exploration, there also emerge some studies that focus on resource construction, to alleviate the lack of labeled training data for supervised pair-wise opinion extraction. Dai and Song (2019) proposed an algorithm to automatically annotate raw data based on the rules derived from dependency parsing and POS tagging results. Li, Chen, Quan, Ling, and Song (2020) proposed a masked sequence-to-sequence method for conditional augmentation of aspect term extraction. However, all of the works presented above focus on the extraction of OT and OW that appear in the original review, but our work does focus on the generation of absent opinion pairs.

Moreover, our work is also related to aspect-based summarization, which refers to generate a paragraph of summary centered around a given aspect. Compared with the normal summarization task, the main challenge of aspect-based summarization is how to effectively encode aspect-specific features. Krishna and Srinivasan (2018) built their model on the pointer network (Vinyals, Fortunato, & Jaitly, 2015), calculating the context vector with aspect features in the decoding stage. Frermann and Klementiev (2019) induced latent document structure jointly with an abstractive summarization objective and trained their models in a scalable synthetic setup. Yang, Qu, Shen, Liu, Zhao, and Zhu (2018) further incorporated sentiment labels into the process of summarization generation.

Our task has some essential differences from aspect-based summarization, e.g., our task has two generation objects namely opinion target and word generation with much shorter but more compact decoding outputs. Inspired by the success of the sequence-to-sequence framework in neural keyphrase generation and summarization (Liu & Lapata, 2019; Meng et al., 2017; Rush et al., 2015; Yuan et al., 2020), we built our model based on the pointer–generator network (See et al., 2017) that generates target–opinion pairs in a correlated hierarchical decoding fashion.

3. Task and dataset

3.1. Task definition for ABPOG

Given a review that consists of a sequence of words $x = \{x_1, x_2, \ldots, x_T\}$ and an aspect category a that belongs to the aspect set $\mathcal{A}$. The goal of the ABPOG task is to generate an opinion pair $\langle$OT, OW$\rangle$, where OT is the opinion target $ot = \{ot_1, ot_2, \ldots, ot_m\}$ that consists of m tokens and OW is the opinion word $ow = \{ow_1, ow_2, \ldots, ow_n\}$ that consists of n tokens.

3.2. Dataset construction

We construct a dataset to train and evaluate our model and task in the following steps (see Fig. 1).

Table 2
An example from the dataset.

r:车子的减震不错,开起来很舒适。本人身高180cm，坐第三排也没问题。车子开起来有很强的推背感，高速上超车毫不费劲(The car has good shock absorption and is very comfortable to drive. I am 180 cm tall, and I have no problem sitting in the third row. I can get a gratifying jump forward when driving the car, so it does not take much effort to overtake at high speed.)
a: space
ot:第三排空间(space of the third row)
ow:够用 (enough)

Table 3
Dataset statistics. Review Length (RL), OT Length (OTL), OW Length (OWL), review where OT is Absent (OTA), review where OW is Absent (OWA).

	Instance	RL	OTL	OWL	OTA	OWA
Train	150,000	128.6	3.1	1.8	71,561	105,625
Dev	10,000	129.1	3.2	1.9	4,831	7,051
Test	10,000	128.3	3.1	1.8	4,760	7,021

Data collection. The raw data are collected from the AUTOHOME,[1] a well-known automotive review site in China. Each review (e.g., the example in Table 1) in the raw data contains an aspect category (e.g., space) and a keyphrase (e.g., 第三排空间够用 (enough space of the third row)), which may not explicitly present in the original text.

Tokenization. To generate opinion target–word pairs, keyphrases are segmented via a Chinese tokenizer with noun and adjective dictionaries in the auto domain.

Opinion target–word pair generation. We use the Stanford Constituency Parser (Manning et al., 2014) to parse tokenized keyphrases into constituent trees. Only the uppermost NP and VP in a constituent tree are used as the OT and OW, respectively. For example, we use the NP "第三排空间" (space of the third row) as the OT and the VP "够用" (enough) as the OW.

Aspect-level rebalance. After generating opinion target–word pairs for each aspect category in each review, we eliminate the aspect categories with extremely fewer data to make the data more balanced. Finally, we obtain six aspect categories, namely *appearance*, *comfort*, *control*, *power*, *interior*, and *space*.

Multi-aspect synthesis. Referring to Frermann and Klementiev (2019), we synthesize multi-aspect reviews by randomly selecting single-aspect reviews from different aspect categories: (1) Randomly select 3~6 aspect categories. (2) Select one review from each aspect category and concatenate them in random order. Note that the selected reviews comment on the same vehicle. (3) The synthetic reviews that are longer than 260 words will be discarded.

Instance construction. Each instance of our dataset can be considered as a quadruple $\langle r, a, ot, ow \rangle$, where r is a review, a is an aspect category, ot and ow are the opinion target and word with regards to the review and aspect. Since a review is multi-aspect, a review is paired with each aspect multiple times. Table 2 shows a constructed instance from the dataset.

Manual double-check. Since the process of data generation is automatic, there exists certain noise. To guarantee the quality of our dataset, we recruit five annotators majoring in Chinese Linguistics to manually check the data. However, considering the size of the data, it is infeasible to check all of them. Instead, the annotators have checked 10,000 instances that are used as the test set of our dataset. The agreement rate for these manually checked instances is 93%.

3.3. Dataset statistics

Since the test set has 10,000 instances, we split the same number of instances from the original data as the development set and employ other 150,000 instances as the training set. Table 3 summarizes the overall data statistics of the training, development, and test sets. As can be seen from the table, the instances of absent OT or OW account for about 47.7% and 79.8%, respectively. As shown in Table 4, although the instance number for each aspect category is not even, there are no categories with extremely few instances. The category *comfort* and *power* contain more instances followed by *space*, *control*, *interior* and *appearance* respectively.

To determine whether the dataset size is appropriate, we refer to keyphrase generation tasks where the length of the keywords is close to our generated target. Table 5 shows the statistics of common keyphrase generation datasets. The first four datasets in the table are based on news or literature abstracts, with one text corresponding to multiple generated objects. The Twitter and Weibo corpora (Wang et al., 2019) are more similar to our task since they also deal with short texts, and the generated objects are more consistent in length. As can be seen, our dataset is larger than the commonly used keyphrase generation dataset so that it can provide enough data for neural models. It is worth noting that we also examined their performance by enlarging the training dataset, but no significant improvement was observed.

[1] https://k.autohome.com.cn/.

Table 4
Instance statistics based on aspect categories.

	appearance	*comfort*	*control*	*power*	*interior*	*space*
Train	18,122	33,233	21,718	34,961	19,992	21,974
Dev	1,218	2,206	1,465	2,301	1,323	1,487
Test	1,182	2,244	1,449	2,275	1,357	1,493

Table 5
Instance statistics based on aspect categories.

Dataset	Train	Dev	Test	Total
Inspec (Hulth, 2003)	1,500	–	500	2,000
Krapivin (Krapivin, Autaeu, & Marchese, 2009)	1,904	–	400	2,304
SemEval (Kim, Medelyan, Kan, & Baldwin, 2010)	144	–	100	244
KP20k (Meng et al., 2017)	20,000	–	20,000	567,830
Twitter (Wang et al., 2019)	35,290	4,411	4,411	44,112
Weibo (Wang et al., 2019)	37,037	4,629	4,629	46,295
StacksExchange (Wang et al., 2019)	39,558	4,944	4,944	49,446

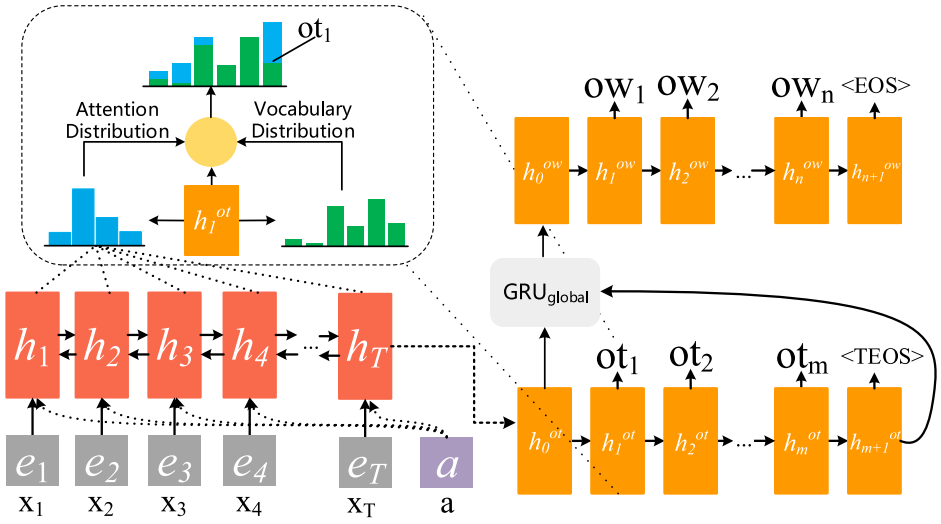

Fig. 2. Overview of our model. x and e (input and embedding), a (aspect), ot (opinion target), ow (opinion word). The opinion word generation on the top-left corner is for ow_1 and based on the pointer–generator network (See et al., 2017).

4. Model framework

4.1. Overview

In this section, we describe the proposed model in detail. Since opinion target–word pairs may exist in the reviews or be absent, we build our model based on the pointer–generator network (See et al., 2017), an encoder–decoder architecture that considers both the generation vocabulary and input text at each decoding step. In addition, we enhance the pointer–generator network to accommodate our task by two methods: (1) We propose a bidirectional aspect-aware GRU (Bi-AGRU) as the encoder to fuse aspect category features. (2) We generate opinion targets (OTs) and words (OWs) jointly by concatenating them with a split token (e.g., OT ⟨TEOS⟩ OW). Besides, we utilize two hierarchical GRUs as the decoder, where a token-level GRU is utilized to provide token features and a global GRU is utilized to provide the decoding status information between OTs and OWs. The framework of our proposed model is illustrated in Fig. 2.

4.2. Encoder side

4.2.1. Aspect-aware GRU

Basic RNN cells such as GRU (Chung, Gülçehre, Cho, & Bengio, 2014) or LSTM (Hochreiter & Schmidhuber, 1997) receive the last hidden state and word embedding as input. Inspired by Liang et al. (2019), we use the **A**spect-aware **GRU** (AGRU) extended from the basic GRU, which receives an additional aspect embedding as input:

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) * \boldsymbol{h}_{t-1} + \boldsymbol{z}_t * \widetilde{\boldsymbol{h}}_t, \tag{1}$$

$$\widetilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_1[\boldsymbol{r}_t * \boldsymbol{h}_{t-1}, \boldsymbol{g}_t * \boldsymbol{e}_t]) + \boldsymbol{l}_t * \boldsymbol{W}_2 \cdot \boldsymbol{e}_t, \tag{2}$$

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z[\boldsymbol{h}_{t-1}, \boldsymbol{e}_t]), \tag{3}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r[\boldsymbol{h}_{t-1}, \boldsymbol{e}_t]), \tag{4}$$

$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}_g[\boldsymbol{a}, \boldsymbol{e}_t]), \tag{5}$$

$$\boldsymbol{l}_t = \sigma(\boldsymbol{W}_l[\boldsymbol{h}_{t-1}, \boldsymbol{a}, \boldsymbol{e}_t]), \tag{6}$$

where $\boldsymbol{z}_t$ and $\boldsymbol{r}_t$ are the update gate and reset gate defined in the basic GRU. $\boldsymbol{e}_t$ and $\boldsymbol{a}$ denote the word and aspect embeddings. Also, we define an aspect gate $\boldsymbol{g}_t$ and a transformation gate $\boldsymbol{l}_t$ to control the non-linear and linear transformation scale of the input $\boldsymbol{e}_t$ with regards to the aspect embedding $\boldsymbol{a}$, which is motivated by Meng and Zhang (2019).

4.2.2. Bi-directional AGRU

Given a review and its word sequence $x = \{x_1, x_2, \ldots, x_T\}$ that maps to an embedding sequence $\boldsymbol{e} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_T\}$, we utilize a bi-directional AGRU layer as the encoder, where the hidden state $\boldsymbol{h}_t$ at the encoding timestep t is calculated as:

$$\overrightarrow{\boldsymbol{h}}_t = \overrightarrow{\mathrm{AGRU}}(\overrightarrow{\boldsymbol{h}}_{t-1}, \boldsymbol{e}_t, \boldsymbol{a}), \tag{7}$$

$$\overleftarrow{\boldsymbol{h}}_t = \overleftarrow{\mathrm{AGRU}}(\overleftarrow{\boldsymbol{h}}_{t+1}, \boldsymbol{e}_t, \boldsymbol{a}), \tag{8}$$

$$\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t], \tag{9}$$

where $\overrightarrow{\mathrm{AGRU}}$ and $\overleftarrow{\mathrm{AGRU}}$ encode the input in the left-to-right and right-to-left orders respectively. Then $\overrightarrow{\boldsymbol{h}}_t$ and $\overleftarrow{\boldsymbol{h}}_t$ are concatenated to $\boldsymbol{h}_t$, which is the final representation for the tth input token.

4.3. Decoder side

The decoding output sequence s of our model is a concatenation of OT and OW words with a split token [TEOS]:

$$s = \{ot, [\mathrm{TEOS}], ow\}, \tag{10}$$

where $ot = \{ot_1, \ldots, ot_m\}$ and $ow = \{ow_1, \ldots, ow_n\}$ indicate the opinion target and word with m and n tokens.

Considering the success of the hierarchical decoding architecture (Chen, Chan, Li, & King, 2020a; Finch, Wang, Utiyama, & Sumita, 2015; Okamoto, Kutsuzawa, Sakaino, & Tsuji, 2018; Zhu, Xue, & Yuan, 2018), we generate the concatenated sequence s with a token-level GRU and a global GRU. Once the OT has been generated, the final state of the token-level GRU will be fed into the global GRU to get the state in order to initialize the start state for OW generation.

4.3.1. Opinion target generation

At each timestep t, we generate the hidden state $\boldsymbol{h}_t^{ot}$ based on the previous hidden state $\boldsymbol{h}_{t-1}^{ot}$ and the embedding of the previous output token $\boldsymbol{ot}_{t-1}$:

$$\boldsymbol{h}_t^{ot} = \mathrm{GRU}_1(\boldsymbol{h}_{t-1}^{ot}, \boldsymbol{ot}_{t-1}), \tag{11}$$

where GRU_1 is the token-level GRU. Then the probability distribution of the output word ot_t is produced by considering both the vocabulary and input tokens. The distribution of the input tokens is calculated via the dot-attention mechanism (Bahdanau, Cho, & Bengio, 2015; Luong, Pham, & Manning, 2015):

$$\boldsymbol{p}^i = \mathrm{softmax}(\boldsymbol{w}^T \, \mathrm{MLP}_1^a([\boldsymbol{h}, \boldsymbol{h}_t^{ot}])), \tag{12}$$

where $\boldsymbol{w}$ is a learnable matrix and MLP_1^a is a multi-layer perceptron corresponding to the aspect type a, which uses the hidden state $\boldsymbol{h}_t^{ot}$ and the encoding token representation $\boldsymbol{h}$ as input. Note that different MLP_1 corresponds to different aspects a for shared-private multi-task learning (Caruana, 1997; Liu, Qiu, & Huang, 2017). The distribution of the vocabulary is calculated as below:

$$\boldsymbol{p}^v = \mathrm{softmax}(\mathrm{MLP}_2^a([\boldsymbol{h}_t^{ot}, \boldsymbol{h}_t^*])), \tag{13}$$

where $\boldsymbol{h}_t^*$ is the input context representation computed by weighted adding all the encoding token representations $\sum_{j=1}^{T} p_j^i \boldsymbol{h}_j$. The final probability distribution can be computed as:

$$\boldsymbol{p} = \mathcal{G}\boldsymbol{p}^v + (1 - \mathcal{G})\boldsymbol{p}^i, \tag{14}$$

$$\mathcal{G} = \sigma(\boldsymbol{W}_{\mathcal{G}}[\boldsymbol{h}_t^*, \boldsymbol{h}_t^{ot}, \boldsymbol{ot}_{t-1}]), \tag{15}$$

where $\boldsymbol{W}_{\mathcal{G}}$ is the weight for the gate, which is used as a soft switch to choose between generating a word from the vocabulary or copying a word from the input sequence. This procedure is repeated so that the words of the opinion target are sequentially emitted until either the token [TEOS] is produced or the maximum output length is reached.

Table 6
Dataset statistics. Review Length (RL), OT Length (OTL), OW Length (OWL), review where OT is Absent (OTA), review where OW is Absent (OWA).

Parameters	Description	Value
Batch	Batch size	64
d_{ew}	Dimension of word embedding	300
d_{ea}	Dimension of word embedding	50
d_h	Dimension of the AGRU of the encoder	150
d_{tg}	Dimension of the token-level GRU of the decoder	300
d_{gg}	Dimension of the global-level GRU of the decoder	300
lr	Initial learning rate	0.001
Dropout	Dropout rate	0.5
Gradient clip	Threshold of Max-Norm Regularization	5

4.3.2. Opinion word generation

As mentioned before, the decoder side consists of a token-level GRU and a global GRU. Under such hierarchical decoding architecture, we utilize the global GRU to initialize the state of the token-level GRU when the OW decoding starts:

$$\boldsymbol{h}_0^{ow} = \mathrm{GRU}_2(\boldsymbol{h}_{m+1}^{ot}, \boldsymbol{h}_0^{ot}), \tag{16}$$

where $\boldsymbol{h}_0^{ow}$ is the initial state of the token-level GRU for OW decoding, GRU_2 is the global GRU, $\boldsymbol{h}_{m+1}^{ot}$ and $\boldsymbol{h}_0^{ot}$ are the hidden states of the token-level GRU in the first and last steps during OT decoding. Compared with the linear decoding structure, the hierarchical decoding structure makes the decoder receive more global information. The generation process for OW is similar to that of OT:

$$\boldsymbol{h}_t^{ow} = \mathrm{GRU}_1(\boldsymbol{h}_{t-1}^{ow}, \boldsymbol{ow}_{t-1}), \tag{17}$$

where $\boldsymbol{h}_t^{ow}$, $\boldsymbol{h}_{t-1}^{ow}$ and $\boldsymbol{ow}_{t-1}$ are the current and previous hidden states, and the embedding of the previous output token, respectively. To train our model, the negative log-likelihood loss is employed:

$$\mathcal{L} = -\frac{1}{m+n+1}\sum_{t=1}^{m+n+1} \mathbf{1}_{s_t} \log \boldsymbol{p}_{s_t}, \tag{18}$$

where s_t denotes the tth step in the output sequence $s = \{ot_1, \ldots, ot_m, [\mathrm{TEOS}], ow_1, \ldots, ow_n\}$, $\mathbf{1}$ is a one-hot vector that represents the gold standard and $\boldsymbol{p}$ is the predicted probability distribution in Eq. (14).

5. Experimental setup

5.1. Evaluation metrics

5.1.1. Exact string match

Since the number of predicted pairs is the same as the number of ground truths, the values of precision, recall and F1 are the same. Therefore, we only calculate recall values concerning opinion ⟨OT, OW⟩ pairs, OT only, and OW only, respectively. For the evaluation of opinion pairs, a predicted ⟨OT, OW⟩ pair is considered to be correct if both OT and OW exactly match the ground truths in the corresponding aspect category.

5.1.2. Semantic similarity match

Since there may be many different terms to express the same attitude, the metric based on exactly matching cannot comprehensively reflect model performances for OW generation. To this end, we also utilize the **R@*k*** metric, where k denotes the threshold of semantic similarity between the predicted opinion word and the ground truth. We use the synonym-lexicon-based approach (Che, Li, & Liu, 2010) to measure the similarity. When the similarity is greater than the predefined threshold k, the predicted opinion word is considered to be correct. Since the expression of OT is usually unambiguous based on our observation, we use R@k only for OW evaluation.

5.2. Hyper-parameters and implementation details

Grid search is employed to determine some optimal values according to the development set. The word and aspect embedding dimensions are randomly initialized according to the uniform distribution in [-0.1,0.1]. The encoder and decoder share the same word embeddings. We tried using a dense layer, a dense layer with a none linear layer or directly copying to obtain the hidden state between the encoder and the decoder. We optimize all model parameters using Adam (Kingma & Ba, 2015) with an initial learning rate 0.001. We use Max-Norm Regularization to clip the gradients for avoiding the gradient explosion. We select the batch size among {8,16,32,64}. Besides, we use dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) after the embedding layer in the encoder and before the MLP layer in the decoder with a probability of 0.5. We list the hyperparameters of our model in Table 6.

Table 7
Recall values on the testing set for OT and OW evaluation separately. *appe.* is short for *appearance*.

Model	OT							OW						
	appe.	*comfort*	*control*	*power*	*interior*	*space*	macro	*appe.*	*comfort*	*control*	*power*	*interior*	*space*	macro
MARS	61.25	91.98	80.88	88.92	90.86	88.95	83.81	50.08	57.84	54.59	56.76	44.13	37.78	50.20
Topic-Oriented	57.28	91.93	81.50	88.97	91.03	87.61	83.05	48.65	58.69	54.87	56.17	44.29	36.72	49.89
Enc-Attn	56.63	87.38	75.89	86.99	88.19	85.11	80.03	45.74	55.25	50.88	54.67	40.35	35.51	47.07
Dec-Attn	55.92	90.95	78.61	88.79	90.64	86.40	81.89	47.80	56.42	51.69	56.26	41.56	36.97	48.45
Source Factors	56.35	91.09	76.81	88.44	88.06	86.81	81.26	46.87	56.64	53.35	56.88	40.83	35.97	49.83
Two-Stage	62.44	91.09	81.57	90.90	89.68	87.81	84.25	49.07	56.28	52.73	57.07	42.96	37.91	49.34
Pipeline	64.97	91.93	81.99	89.10	90.94	**88.81**	84.62	**53.21**	57.09	55.76	57.49	44.22	38.18	50.99
Joint	**67.60**	**92.56**	83.51	**91.08**	91.01	87.74	**85.58**	51.52	54.72	50.86	**57.67**	44.66	34.49	48.99
Joint+hrd	61.08	92.20	**83.78**	89.98	**91.75**	88.68	84.58	51.61	**59.58**	**55.90**	57.19	**45.54**	**38.45**	**51.38**

5.3. Baselines and our model variations

5.3.1. Baselines

Since 47.7% of OT and 79.8% of OW do not appear in the original review, the extraction-based approach would perform very poorly on the dataset, so we only consider the generation-based approaches.

As ABOPG task received extra input as a constraint of the generation, we consider the start-of-the-art methods for aspect-based summarization task and other related tasks listed below as the baselines:

- **MARS** (Yang et al., 2018) propose a mutual attention mechanism to interactively learns the representations of context words and aspect acted as an encoder. The learned aspect representations are incorporated into the decoder to generate aspect-aware review summaries via an attention fusion network.
- **Topic-Oriented** (Krishna & Srinivasan, 2018) uses the original PG-net, modifies its input by treating the target aspect as additional information (factor) and concatenates the aspect embedding as additional information (factor) to the embedding of each word.
- **Enc-Attn** (Frermann & Klementiev, 2019) utilizes pointer generator as the backbone and fuses aspect features in the encoding stage by adding an attention mechanism to the encoder. They calculate an aspect-specific weight for each token representation and scale each latent representation independently by passing the weight through a sigmoid function. This method utilizes the pointer generator as backbone, too.
- **Dec-Attn** (Frermann & Klementiev, 2019) integrates aspect features during the decoding phase. They learn separate attention weights and biases for each possible input aspect, and use the parameters specific to target-aspect during decoding. This method also utilizes the pointer generator as backbone.
- **Source-Factors** (Frermann & Klementiev, 2019) is similar to **Topic-Oriented** and also concatenates the aspect embedding to each input word.
- **Two-Stage** (Hayashi et al., 2021) labels sentences with aspect and then group sentences with the given aspect in order of occurrence in cited references to form a chunked paragraph that discusses the same aspect, which become the input to a summarization model, and are then feed into the aspect-specific pointer generator.

All the baselines are based on the pointer–generator network, and we use different MLP in the decoder corresponds to the input aspect a like Equation (12) (13) to ensure fairness of comparison.

5.3.2. Model variations

To investigate the impacts of different parts in our model, we also design several variations. **Pipeline** generates OT and OW pairs separately via two token-level decoders. **Joint** generates OT and OW pairs jointly but does not use the global GRU. **Joint+hrd** generates OT and OW pairs jointly and uses both the token-level and global GRUs to form the hierarchical decoding architecture.

6. Results and analysis

6.1. Separately evaluation for OT and OW

As shown in Table 7, we analyze the results of OT and OW generation separately to gain a fine-grained understanding of the performance of the models. Overall, the recall values of OW generation are generally lower than those of OT generation with regards to most models, indicating that OW generation is more challenging than OT generation and thus constrains the overall performance.

Moreover, the *Joint* model has better performances than the *Pipeline* model in OT generation, while the results for OW are reversed. This may be because the *Joint* model can optimize the parameters associated with OT and OW simultaneously. However, since the decoding sequence for the *Joint* model is longer than that of the *Pipeline*, errors are more likely to occur when generating the OW at the back.

After adding the *hrd* module, the recall value of *Joint+hrd* drops by 1% for OT, and roses by 2.39% for OW. This may be due to the fact that the hierarchical decoder considers more whether the two elements match rather than the impact of individual elements.

Table 8
Recall values for ⟨OT, OW⟩ pair evaluation.

Model	*appearance*	*comfort*	*control*	*power*	*interior*	*space*	macro
MARS	46.38	54.21	47.65	53.29	41.89	35.24	46.44
Topic-Oriented	46.38	54.21	47.65	53.29	40.72	34.24	46.08
Enc-Attn	44.89	52.27	45.36	52.33	36.54	33.42	44.13
Dec-Attn	45.94	53.83	45.82	52.57	39.79	34.49	45.41
Source Factors	45.18	54.14	46.31	53.32	39.50	34.56	45.50
Two-Stage	43.74	53.88	46.31	**55.25**	40.09	36.04	45.89
Pipeline	43.74	53.57	47.69	52.09	41.12	35.30	45.58
Joint	45.60	52.54	46.79	54.99	43.04	33.09	46.01
Joint+hrd	**47.29**	**56.68**	**50.38**	53.23	**43.92**	**36.70**	**48.04**

Table 9
Recall values for ⟨OT, OW⟩ pair evaluation based on semantic similarity.

Model	R	R@1	R@0.9	R@0.8	R@0.7
MARS	46.44	55.91	56.27	57.62	58.71
Topic-Oriented	46.08	55.41	56.91	57.60	58.61
Enc-Attn	44.13	53.33	53.74	54.64	55.91
Dec-Attn	45.41	54.77	55.27	56.42	57.63
Source Factors	45.50	54.33	54.73	56.04	57.42
Two-Stage	45.89	54.94	55.47	56.16	57.91
Pipeline	45.58	55.16	55.60	56.54	57.76
Joint	46.01	56.19	56.54	57.05	58.10
Joint+hrd	**48.04**	**57.55**	**57.86**	**58.38**	**59.40**

Meanwhile, there are fewer steps to generate OW through global GRU, so the performance of OW generation is better. We further analyze 47 *hrd* on the elements and the whole pair in the next section.

6.2. Evaluation for opinion target–word pairs

We report the performance of our model for opinion pair generation in Table 8. The results for the *Dec-Attn* and *Source Factors* models are not significantly different. Our final model *Joint+hrd* outperforms the best performing baseline, namely *MARS*, by 1.6%. Specifically, compared to *Pipeline*, jointly modeling the generation of OT and OW, the recall of *Joint* model rises by 0.43%. After utilizing the hierarchical decoding architecture, our *Joint+hrd* model achieves the best performance with a macro recall 48.04%. Moreover, the *Joint+hrd* model also achieves the highest recall values across almost all aspect categories except *power*. In terms of per-category performance, our model generally performs better for the categories that account for higher proportions in the dataset. For example, all models perform well for the aspect category *power*.

Notice that the *Joint+hrd* model did not get the highest recall values on the test sets for OT and OW of aspect appearance in Table 7, but achieved the highest recall values for opinion pair in Table 8. This is because *Joint+hrd* model can generate two more matching elements. For example, *Pipeline* achieves higher recall values on both OT and OW but lower recall values on the opinion pair than *Joint+hrd*. It indicates that the OT and OW correctly predicted by pipeline are not in the same pair, i.e., only one element in the same pair is correctly predicted. Generating two elements that match each other is one of the challenges of the ABOPG task. Based on this motivation, we proposed the hierarchical decoder to take both completely generated OT and context during OW generation.

6.3. Evaluation based on semantic similarity

Besides strict string-matching evaluation, we also perform the evaluation for OW generation based on semantic similarity, where the value k in R@k indicates the semantic similarity. As shown in Table 9, the R@k values for all the models are significantly higher than the recall values without using the semantic similarity (Column 2). This demonstrates that the strict string-matching evaluation cannot reflect the model performance comprehensively, since opinion words are various and different opinion words may express the same meaning. In addition, when the similarity k keeps decreasing, recall values do not change much, indicating that the evaluation method is stable and not very sensitive to the similarity.

6.4. Evaluating the "present–absent" OT and OW

Recall that opinion targets or words may not occur in the review text of some instances in our dataset. We call such opinion targets or words and the related instances as "absent" items. In this section, we compare the results of absent and present items in Tables 11 and 10, respectively. As we can see, our *Joint+hrd* model outperforms all the baselines on all the metrics. Comparing the results of present and absent items, the performance of the present items is much higher than that of the absent. This demonstrates that it is more difficult for the OT or OW generation for the instances where OT or OW is absent in the input text, and the generation of absent items is crucial to improve the overall performance.

Table 10
Recall values of the Present.

Model	OT	OW			Pair-level		
	R	R	R@1	R@0.7	R	R@1	R@0.7
MARS	90.23	56.76	67.78	72.65	57.45	65.88	69.76
Topic-Oriented	90.31	56.05	67.69	71.85	57.48	66.46	68.42
Enc-Attn	89.01	55.46	65.23	67.81	56.54	64.52	65.02
Dec-Attn	89.98	56.06	66.48	70.07	57.73	65.33	68.75
Source Factors	89.64	56.31	66.35	70.24	57.69	66.45	69.05
Two-Stage	89.45	55.37	65.05	67.40	57.72	66.19	69.88
Pipeline	90.87	56.82	68.38	72.59	56.44	65.19	68.54
Joint	90.98	57.68	68.53	71.73	58.26	66.75	69.06
Joint+hrd	**91.02**	**59.36**	**70.75**	**74.26**	**59.57**	**68.23**	**70.26**

Table 11
Recall values of the Absent.

Model	OT	OW			Pair-level		
	R	R	R@1	R@0.7	R	R@1	R@0.7
MARS	78.60	43.65	53.73	58.86	41.59	51.23	53.94
Topic-Oriented	77.10	43.86	53.71	58.74	42.10	51.67	53.97
Enc-Attn	75.05	41.13	50.97	56.16	40.03	49.54	52.21
Dec-Attn	76.35	42.00	52.05	57.70	41.29	50.91	53.76
Source Factors	75.71	42.21	52.35	57.43	41.46	50.28	53.35
Two-Stage	79.45	43.55	54.37	58.40	40.95	50.96	53.45
Pipeline	80.55	43.53	54.54	58.77	41.90	51.53	53.95
Joint	**81.16**	43.08	55.15	59.15	41.97	52.47	54.35
Joint+hrd	80.24	**45.80**	**56.14**	**60.44**	**44.16**	**53.88**	**55.74**

Table 12
Ablation study of the Joint+hrd model. Models are evaluated on recall values for ⟨OT, OW⟩ pair based on exact string match (denoted as R) and semantic similarity (denoted as R@k).

Model	R	R@1	R@0.9	R@0.8	R@0.7
Joint+hrd	**48.04**	**57.55**	**57.86**	**58.38**	**59.40**
-AGRU	44.74	53.78	54.15	55.43	56.84
-Hierarchical Decoder	40.99	49.81	50.18	51.28	52.41
-Pointer Mechanism	47.74	57.31	57.47	58.06	59.38

6.5. Ablation study

To study the contribution of each component in the Joint+hrd model, we ran an ablation study on the test set as shown in Table 12. We find that: (1) The AGRU contributes 3.3 Recall, which shows the capabilities of AGRU. (2) Recall drops by 7.05 when we remove the hierarchical decoder structure and only use one local GRU for decoding, which proves that global information is critical in generating an opinion pair. (3) Pointer mechanism only contributes 0.3 Recall, which shows that distribution of the input tokens is not as important as we expected. This is probably because there are almost no OOV words in the generated content.

6.6. Case study

To analyze the effects of jointly decoding ⟨OT, OW⟩ pairs and hierarchical decoding architecture, we select two cases that are predicted by the *Pipeline*, *Joint* and *Joint+hrd* models, as shown in Table 13. The *Pipeline* model suffers from error propagation. For example, it generates wrong OWs (e.g., good) after predicting wrong OTs (e.g., brake pedal) for the aspect *control* in the first case. By contrast, the *Joint* model correctly predicts the OT (e.g., force to apply the brakes) even though the OW is wrongly predicted (e.g., 合) due to the joint optimization decoding method. However, the *Joint* model may be apt to make mistakes during generating OWs (e.g., 合). In contrast, the *Joint+hrd* model can avoid the above problem by utilizing the global GRU for hierarchical decoding. In addition, the *Pipeline* model generates "brake pedal" and "good" as the OT and OW for the aspect *control* in the first case, which are semantically similar to the ground truths. Therefore, this example shows that it is necessary to introduce semantic-similarity-based evaluation.

6.7. Interpretability analysis

To explore whether our proposed method is interpretable, we visualize the sum of the attention weights (Eq. (12)) received by each word over all decoding steps, as shown in Fig. 3. The degree of the color is used to indicate the size of the weight (the

Table 13
Case Study. Aspect-specific OTs and OWs generated by different methods for reviews from test set. Items that were correctly and incorrectly predicted are highlighted in blue and red, respectively.

Review	Aspect	Pipeline		Joint		Joint+hrd	
		OT	OW	OT	OW	OT	OW
1.新车还是有异味，自己买了一些竹炭。刹车一踩就停，开起来挺舒服的。后斗加了卷帘盖装载能力不是普通SUV能比拟的。 The new car still smells bad, I bought some bamboo charcoal. The brakes stop as soon as you step on them, and it's quite comfortable to drive. The rear bucket has a roll-up cover so that the loading capacity is not that of a normal SUV.	*interior*	异味 odor	浓 strong	异味 odor	浓 strong	异味 odor	浓 strong
	control	刹车踏板 brake pedal	不错 good	刹车踏板轻重 force to apply the brakes	合 ✗	刹车踏板轻重 force to apply the brakes	合适 right
	space	储物空间设计 design of storage space	小 small	储物空间设计 design of storage space	合 ✗	储物空间设计 design of storage space	合理 reasonable
2.控制台用料感觉一般，硬塑料太多。市区开动力输出还不错，时速上了120KM/H后加速能力较弱，日常用足矣，不过刹车偏软。 The material used in the control feels general, too much hard plastic. The power output is not bad when driving in the city, and the acceleration ability is weak after 120KM/H, which is sufficient for daily use, but the brake is soft.	*interior*	中控台材质 material of the console	差 poor	中控台材质 material of the console	慢 slow	中控台材质 material of the console	一般 general
	control	刹车 brake	软 soft	刹车 brake	软 soft	刹车 brake	软 soft
	power	市区路速 ✗	慢 slow	市区路段动力 power on city roads	足够 adequate	市区路段动力 power on city roads	足够 adequate

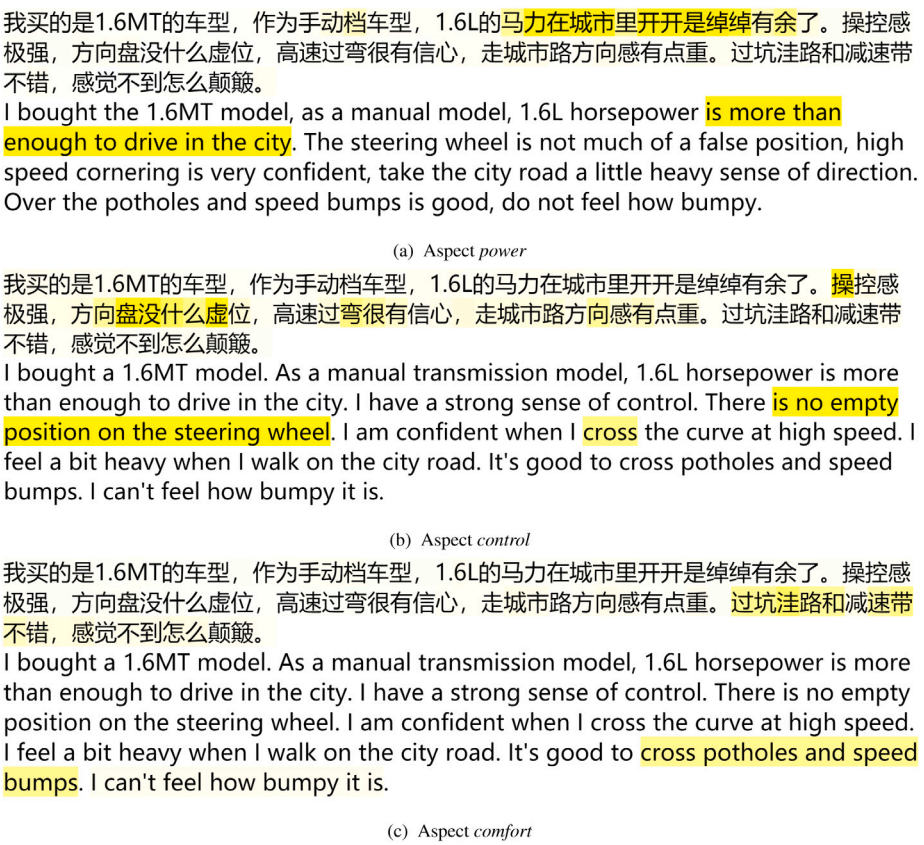

Fig. 3. Attention weights accumulated over all decoding steps. The darker the color is, the larger the weight is. Due to the differences in Chinese and English grammar, it is not possible to translate word by word, only the snippets with larger weight are identified in the English translation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

darker the larger). As we can see, the attention distribution varies according to the given aspect of a review. Moreover, the words or snippets that are closely related to a given aspect category receive more attention. For example, given the aspect *power*, the text snippet "马力在城市里开绰绰有余" (Horsepower is more than enough in the city) is highlighted with more yellow, which demonstrates that it obtains more attention as shown in Fig. 3(a). Fig. 3(b) shows the weighted sum for the aspect *control*, in which we can see that the text snippet "方向盘没什么虚位" (There is no clearance in steering wheeling) and "过弯" (turn the corner) get more attention. As shown in Fig. 3(c), the text snippet "过减速带和坑洼路不错" (It's good to pass the speed bumps and potholes) and "还感觉不到怎么颠簸" (I don't feel any bumps) have the darkest color.

Table 14
Comparison of ABPOG and related tasks.

Task	ABPOG	ABS	OPE
Input	Review	News	Review
Aspect as Input	Yes	Yes	No
Output	⟨OT, OW⟩	Paragraph	⟨OT, OW⟩
Features	The desired OTs and OWs may not present in the original text.	ABS aims to summarize and refine the original text according to the given aspect.	OPE can only obtain the OTs and OWs that appear in the original text.

7. Discussion

To explain the difference between our work and the previous works, we elaborate from four aspects: task, dataset, model, and application.

In contrast to previous studies, our study focuses on the generation of opinion pairs that do not appear in the original text. Table 14 lists the features of ABPOG and other related tasks include aspect-based summarization (ABS) (Frermann & Klementiev, 2019; Kunneman et al., 2018) and opinion pair extraction (OPE) (Chen et al., 2020b; Wan et al., 2020). Compared with ABS task, which distills and streamlines the original text, ABPOG can obtain more fine-grained information, e.g. opinion target–word pairs from reviews. Compared with OPE task, ABPOG can obtain opinion target–word pairs which not appear in the original text, and generate aspect-specific content based on the input aspect.

Existing corpora of fine-grained opinion analysis, especially the OPE task, usually consist of the original texts, OTs, OWs, and the relationships among them. Due to the limitation of extraction-based methods, both OTs and OWs all appear in the original text. In the dataset we constructed, about 47.7% OT and 79.8% OW did not appear in the original text. Also, there is an extra aspect field since ABPOG is an aspect-based generation task.

We apply AGRU for the first time to extract aspect-specific features on an aspect-based text generation task. We use two hierarchical decoders to strengthen the correlation between two generation targets. The experimental results on the constructed dataset provided the effectiveness of our method built on the pointer network. We further propose a semantic-based evaluation approach to enable a more rational evaluation of this task. The case study and model visualization demonstrate that our model can generate reasonable results and meanwhile hold certain interpretability.

In terms of application, the generation of ⟨OT, OW⟩ pairs expressed in the text is a supplement to the OPE task and is informative to potential consumers and businesses. Secondly, in the review-based recommendation systems, ⟨OT, OW⟩ pairs generated according to specific aspects can provide information to meet the needs of users more precisely.

8. Conclusion

In this paper, we explore a novel task, namely Aspect-Based Pair-wise Opinion Generation (ABPOG) and build a multi-aspect dataset based on Chinese automotive reviews. We build our model based on the pointer–generator network to perform the task. Moreover, we exploit an aspect-aware GRU to extract aspect-specific features, and utilize a hierarchical decoding approach to decode opinion target–word pairs. Extensive experiments show that our model outperforms strong baselines. In the future, we will explore how our task and model facilitate aspect-based sentiment analysis that opinion targets and words are absent.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61772378), the National Key Research and Development Program of China (No. 2017YFC1200500), the Research Foundation of Ministry of Education of China (No. 18JZD015), the Major Projects of the National Social Science Foundation of China (No. 11&ZD189).

References

Akhtar, M. S., Garg, T., & Ekbal, A. (2020). Multi-task learning for aspect term extraction and aspect sentiment classification. *Neurocomputing, 398*, 247–256.

Bagheri, A. (2019). Integrating word status for joint detection of sentiment and aspect in reviews. *Journal of Information Science, 45*(6), 736–755.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd international conference on learning representations*.

Caruana, R. (1997). Multitask learning. *Machine Learning, 28*(1), 41–75.

Cavalcanti, D. C., & Prudêncio, R. B. C. (2017). Unsupervised aspect term extraction in online drugs reviews. In *Proceedings of the thirtieth international florida artificial intelligence research society conference* (pp. 38–43).

Che, W., Li, Z., & Liu, T. (2010). LTP: A Chinese language technology platform. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 13–16).

Chen, W., Chan, H. P., Li, P., & King, I. (2020). Exclusive hierarchical decoding for deep keyphrase generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1095–1105).

Chen, S., Liu, J., Wang, Y., Zhang, W., & Chi, Z. (2020). Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6515–6524).

Chung, J., Gülçehre, Ç., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555.

Dai, H., & Song, Y. (2019). Neural aspect and opinion term extraction with mined rules as weak supervision. In *Proceedings of the 57th conference of the association for computational linguistics* (pp. 5268–5277).

Diaz, G. O., Zhang, X., & Ng, V. (2020). Aspect-based sentiment analysis as fine-grained opinion mining. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6804–6811).

Fan, F., Feng, Y., & Zhao, D. (2018). Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3433–3442).

Fan, Z., Wu, Z., Dai, X., Huang, S., & Chen, J. (2019). Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics* (pp. 2509–2518).

Finch, A. M., Wang, X., Utiyama, M., & Sumita, E. (2015). Hierarchical phrase-based stream decoding. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1089–1094).

Frermann, L., & Klementiev, A. (2019). Inducing document structure for aspect-based summarization. In *Proceedings of the 57th conference of the association for computational linguistics* (pp. 6263–6273).

Hayashi, H., Budania, P., Wang, P., Ackerson, C., Neervannan, R., & Neubig, G. (2021). WikiAsp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, *9*, 211–225.

He, R., Lee, W. S., Ng, H. T., & Dahlmeier, D. (2017). An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 388–397).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 216–223).

Kim, S. N., Medelyan, O., Kan, M., & Baldwin, T. (2010). SemEval-2010 task 5 : automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 21–26).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd international conference on learning representations*.

Krapivin, M., Autaeu, A., & Marchese, M. (2009). *Large dataset for keyphrases extraction: Technical report*, University of Trento.

Krishna, K., & Srinivasan, B. V. (2018). Generating topic-oriented summaries using neural attention. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics* (pp. 1697–1705).

Kunneman, F., Wubben, S., van den Bosch, A., & Krahmer, E. (2018). Aspect-based summarization of pros and cons in unstructured product reviews. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2219–2229).

Kushwaha, A., & Chaudhary, S. (2017). Review highlights: opinion mining on reviews: a hybrid model for rule selection in aspect extraction. In *Proceedings of the 1st international conference on internet of things and machine learning* (pp. 1–6).

Li, X., Bing, L., Lam, W., & Shi, B. (2018). Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 946–956).

Li, X., Bing, L., Li, P., Lam, W., & Yang, Z. (2018). Aspect term extraction with history attention and selective transformation. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence* (pp. 4194–4200).

Li, K., Chen, C., Quan, X., Ling, Q., & Song, Y. (2020). Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7056–7066).

Li, S., Zhou, L., & Li, Y. (2015). Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures. *Information Processing and Management*, *51*(1), 58–67.

Liang, Y., Meng, F., Zhang, J., Xu, J., Chen, Y., & Zhou, J. (2019). A novel aspect-guided deep transition model 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* (pp. 5568–5579).

Liao, M., Li, J., Zhang, H., Wang, L., Wu, X., & Wong, K. (2019). Coupling global and local context for unsupervised aspect extraction. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 4578–4588).

Liu, P., Joty, S. R., & Meng, H. M. (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* (pp. 1433–1443).

Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3728–3738).

Liu, P., Qiu, X., & Huang, X. (2017). Adversarial multi-task learning for text classification. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 1–10).

Luo, Z., Huang, S., & Zhu, K. Q. (2019). Knowledge empowered prominent aspect extraction from product reviews. *Information Processing and Management*, *56*(3), 408–423.

Luo, H., Li, T., Liu, B., Wang, B., & Unger, H. (2019). Improving aspect term extraction with bidirectional dependency tree representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(7), 1201–1212.

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1412–1421).

Ma, D., Li, S., Wu, F., Xie, X., & Wang, H. (2019). Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th conference of the association for computational linguistics* (pp. 3538–3547).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 55–60).

Mehta, P., & Majumder, P. (2018). Effective aggregation of various summarization techniques. *Information Processing and Management*, *54*(2), 145–158.

Meng, F., & Zhang, J. (2019). DTMT: A novel deep transition architecture for neural machine translation. In *Proceedings of the thirty-third AAAI conference on artificial intelligence* (pp. 224–231).

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., & Chi, Y. (2017). Deep keyphrase generation. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 582–592).

Meskele, D., & Frasincar, F. (2020). ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. *Information Processing and Management*, *57*(3), Article 102211.

Mukherjee, R., Peruri, H. C., Vishnu, U., Goyal, P., Bhattacharya, S., & Ganguly, N. (2020). Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1825–1828).

Okamoto, T., Kutsuzawa, K., Sakaino, S., & Tsuji, T. (2018). Trajectory planning by variable length chunk of sequence-to-sequence using hierarchical decoder. In *IEEE 15th international workshop on advanced motion control* (pp. 209–214).

Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., & Si, L. (2020). Knowing what, how and why: a near complete solution for aspect-based sentiment analysis. In *The thirty-fourth AAAI conference on artificial intelligence* (pp. 8600–8607).

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 379–389).

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: summarization with pointer-generator networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 1073–1083).

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Tang, D., Qin, B., & Liu, T. (2016). Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 214–224).

Tao, J., & Zhou, L. (2020). A weakly supervised WordNet-guided deep learning approach to extracting aspect terms from online reviews. *ACM Transactions on Management Information Systems*, *11*(3), 13:1–13:22.

Varghese, R., & Jayasree, M. (2013). Aspect based Sentiment Analysis using support vector machine classifier. In *International conference on advances in computing, communications and informatics* (pp. 1581–1586).

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. In *Proceedings of the advances in neural information processing systems 28: annual conference on neural information processing systems* (pp. 2692–2700).

Wan, H., Yang, Y., Du, J., Liu, Y., Qi, K., & Pan, J. Z. (2020). Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the thirty-fourth AAAI conference on artificial intelligence* (pp. 9122–9129).

Wang, Y., Li, J., Chan, H. P., King, I., Lyu, M. R., & Shi, S. (2019). Topic-aware neural keyphrase generation for social media language. In *Proceedings of the 57th conference of the association for computational linguistics* (pp. 2516–2526).

Wang, W., & Pan, S. J. (2019). Transferable interactive memory network for domain adaptation in fine-grained opinion extraction. In *Proceedings of the thirty-third AAAI conference on artificial intelligence* (pp. 7192–7199).

Wang, W., Pan, S. J., Dahlmeier, D., & Xiao, X. (2016). Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 616–626).

Wang, W., Pan, S. J., Dahlmeier, D., & Xiao, X. (2017). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the thirty-first AAAI conference on artificial intelligence* (pp. 3316–3322).

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*(2–3), 165–210.

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2018). Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 592–598).

Yang, M., Qu, Q., Shen, Y., Liu, Q., Zhao, W., & Zhu, J. (2018). Aspect and sentiment aware abstractive review summarization. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1110–1120).

Yin, D., Wu, X., & Chang, B. (2020). Interactive neural network: leveraging part-of-speech window for aspect term extraction (student abstract). In *Proceedings of the thirty-fourth AAAI conference on artificial intelligence* (pp. 13977–13978).

Yu, J., Jiang, J., & Xia, R. (2019). Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(1), 168–177.

Yuan, X., Wang, T., Meng, R., Thaker, K., Brusilovsky, P., He, D., et al. (2020). One size does not fit all: generating and evaluating variable number of keyphrases. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7961–7975).

Zhao, H., Huang, L., Zhang, R., Lu, Q., & Xue, H. (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* (pp. 3239–3248).

Zhu, Z., Xue, Z., & Yuan, Z. (2018). Automatic graphics program generation using attention-based hierarchical decoder. In *Proceedings of the 14th asian conference on computer vision (vol. 11366)* (pp. 181–196).