Pre-trained Personalized Review Summarization with Effective Salience Estimation

Hongyan Xu[1,3], Hongtao Liu[2]*, Zhepeng Lv[2], Qing Yang[2], Wenjun Wang[1]†
[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] Du Xiaoman Financial, Beijing, China
[3] Georgia Tech Shenzhen Institute, Tianjin University, Guangdong, China
[1]{hongyanxu, wjwang}@tju.edu.cn
[2]{liuhongtao01, lvzhepeng, yangqing}@duxiaoman.com

Abstract

Personalized review summarization in recommender systems is a challenging task of generating condensed summaries for product reviews while preserving the salient content of reviews. Recently, Pretrained Language Models (PLMs) have become a new paradigm in text generation for the strong ability of natural language comprehension. However, it is nontrivial to apply PLMs in personalized review summarization directly since there are rich personalized information (e.g., user preferences and product characteristics) to be considered, which is crucial to the salience estimation of input review. In this paper, we propose a pre-trained personalized review summarization method, which aims to effectively incorporate the personalized information of users and products into the salience estimation of the input reviews. We design a personalized encoder that could identify the salient contents of the input sequence by jointly considering the semantic and personalized information respectively (i.e., ratings, user and product IDs, and linguistic features), yielding personalized representations for the input reviews and history summaries separately. Moreover, we design an interactive information selection mechanism that further identifies the salient contents of the input reviews and selects relative information from the history summaries. The results on real-world datasets show that our method performs better than the state-of-the-art baselines and could generate more readable summaries.

1 Introduction

Personalized review summarization aims to generate brief summaries for product reviews and preserves the main contents of input reviews. It could help users to have a full insight of products quickly and make accurate purchase decisions, hence it is an important task in recommender systems and attracts more and more attention recently (Ganesan et al., 2010; Di Fabbrizio et al., 2014; Xiong and Litman, 2014; Gerani et al., 2014; Li et al., 2017; Chan et al., 2020). Different from the traditional summary generation task, reviews are generally coupled with a lot of essential information about users and products. Hence, some recent works propose to generate personalized summaries for reviews by considering user preferences and product characteristics (Li et al., 2017; Yang et al., 2018; Li et al., 2019c,a,b; Chan et al., 2020; Xu et al., 2021). For example, Li et al. (2019a) propose a user-aware encoder-decoder framework that considers the user preferences in the encoder to select important information and incorporate the preferences and writing style of users into the decoder to generate personalized summaries.

Review: I got this seat to replace the deteriorating saddle on my 10-year-old trek navigator 200, while the new saddle certainly has less padding than the old one. It's still surprising comfortable. I like how it's smaller size and profile makes it easier to get on off the bike.
Summary: slimming step down for comfort saddles.

Historical Summaries
1) comfortable at a great price
2) good replacement saddle
3) hard but narrow
4) well at least it's a seat
5) comfy mountain ride

Figure 1: An example of a product review and the corresponding summary. Different history summaries mention different significant features of the given review and we mark them in different colors.

Recently, pre-trained language models (PLMs) have achieved notable improvements in various text generation tasks including abstractive summarization (Amplayo et al., 2021; Li et al., 2022). However, the exploration of the pre-training paradigm in review summarization is quite preliminary since applying PLMs to review summarization directly is nontrivial. The existing PLMs ignore the personalized information about users and products which is crucial to generate personalized summaries for

*Hongtao Liu is the co-first author.
†Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2023, pages 10743–10754

product reviews. First, the salience of the input reviews is not only dependent on the semantic information but also influenced by the personalized information. For example, different users have different preferences towards product characteristics, in Figure 1, the user of history summary 3) is interested in "quality" while the user of history summary 1) focuses more on "price". Therefore, it is necessary to consider user preferences and product characteristics when calculating the salience for the contents of the input review. Second, history summaries of users and products convey rich text descriptions, which could not only be used to strengthen salient information identification of the input reviews but also be fed into the generation process as additional input. For example, some important aspects (e.g., "price", "size", etc.) of products are usually mentioned by different users, hence history summaries might contribute to the salience calculation of the input reviews.

Therefore, it is essential to fine-tune PLMs effectively to make the model could generate personalized summaries by jointly considering semantic information of input reviews and the essential attributes of users and products. However, the existing PLMs focus on the text content, and it is challenging to integrate the various kinds of auxiliary information into PLMs selectively and effectively. In this paper, we propose an encoder-decoder **P**re-trained **P**ersonalized **R**eview **S**ummarization method with effective salience estimation of the rich input information, named *PPRS*. Specifically, we design two kinds of mechanisms to leverage the user and product information to identify the salient contents of the input reviews and history summaries, making the model could focus on the more relevant content towards the current summary generation.

First, we propose a personalized encoder that learns representations for the input reviews and each history summary separately. Considering the user and product IDs indicate their intrinsic characteristics, the personalized encoder aligns each word and the corresponding user, product, and rating. In this way, the salience of words of the input reviews is influenced by both semantic and personalized information, yielding personalized representations. Furthermore, we observe that linguistic features are generally associated with the user opinions and product characteristics, such as users utilizing adjectives to represent their sentiment (e.g., "comfortable", "good" in Figure 1) and aspects of products are usually nouns (e.g., "price", "size" in Figure 1). Therefore, we also aggregate the part-of-speech feature into the personalized encoder to identify the salient content of the input reviews more accurately.

Second, we propose an interactive information selection mechanism that interactively models the input reviews and history summaries to learn more comprehensive representations. On the one hand, considering history summaries are usually noisy and redundant, we select the relevant information from history summaries by calculating the semantic relatedness between history summaries and the input reviews. On the other hand, we learn the history summaries-aware salience for the input reviews by calculating the semantic similarity between words of the input reviews and history summaries. Finally, we combine the input reviews and history summaries as the input of the decoder to generate coherent and personalized summaries.

The main contribution is threefold: **(1)** we propose a PLM-based personalized review summarization method that conducts salience estimation by jointly considering the user and product information. **(2)** we design two mechanisms to incorporate the personalized information into the generation process, i.e., the personalized encoder and an interactive information selection module. **(3)** we conduct extensive experiments and the results show that our method outperforms competitive baselines.

2 PROPOSED METHOD

In this paper, we conduct a review summarization based on Transformer (Vaswani et al., 2017) encoder-decoder architecture initialized with T5 (Raffel et al., 2020). In this section, we first introduce the problem formulation and then describe our method from two aspects: the personalized encoder module as shown in Figure 2 and the decoder with interactive information selection as shown in Figure 3.

2.1 Problem Formulation

Given review X and the corresponding personalized information $A = \{u, v, r, S\}$, our method aims to generate personalized summary $\hat{Y}$, where u is the user ID, v is the product ID, r is the rating given by u to v, and S is the set of history summaries. Especially, the history summary set $S = \{S_1, \cdots, S_M\}$ is constructed by collecting M

summaries of the corresponding user u and product v. In this paper, the input review is represented as $X = \{w_1, w_2, \cdots, w_L\}$ where w_i is the i-th word and L is the number of words. Besides, the generated and reference summaries are denoted as $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_{\hat{T}}\}$ and $Y = \{y_1, y_2, \cdots, y_T\}$ respectively, where $\hat{T}$ and T is the number of words of generated and reference summary respectively.

2.2 Personalized Encoder

In this section, we introduce the personalized encoder which learns a comprehension representation by jointly considering the semantic information and various attributes of the corresponding user and product. As shown in Figure 2, the input reviews and each history summary are encoded separately, hence we take the input review X as an example to introduce the personalized encoder.

In contrast to traditional summarization, we need to consider user preferences, writing style, and product characteristics in review summarization, in order to select the salient content of the input reviews accurately for different users/products. Besides, the rating reflects the sentiment tendency of users toward current products, and hence could be utilized to identify the useful content of input reviews. Therefore, we propose to align each word to rating, user, and product IDs, aiming to learn more comprehension representation for the input review. Additionally, linguistic features are important to identify the salient content of input texts, such as adjectives typically reflect users' opinions (e.g., "good", "bad", etc), and nouns generally reflect product characteristics (e.g., "speed", "price", etc). Therefore, we propose to incorporate the part-of-speech feature into the encoder by considering the part-of-speech of each word in the embedding layer. Finally, the embedding for each word e is denoted as follows:

$$\mathbf{e} = \mathbf{e}_t + \mathbf{e}_p + \mathbf{e}_{pos} + \mathbf{e}_r + \mathbf{e}_u + \mathbf{e}_v, \quad (1)$$

Where $\mathbf{e}_t$ and $\mathbf{e}_p \in \mathcal{R}^{d_e}$, $\mathbf{e}_{pos}$, $\mathbf{e}_r$, $\mathbf{e}_u$, and $\mathbf{e}_v \in \mathcal{R}^{d_a}$ are token, position, part-of-speech, rating, user ID and product ID embedding respectively. Subsequently, the input review is fed into Transformer encoder layers (Vaswani et al., 2017). Specifically, the encoder consists of stacked identical layers, where each layer has two sub-layers: a self-attention network and a fully connected feed-forward network. The encoder could learn a comprehensive representation of the input long sequence by jointly considering semantic and personalized features during the calculation in each encoder layer.

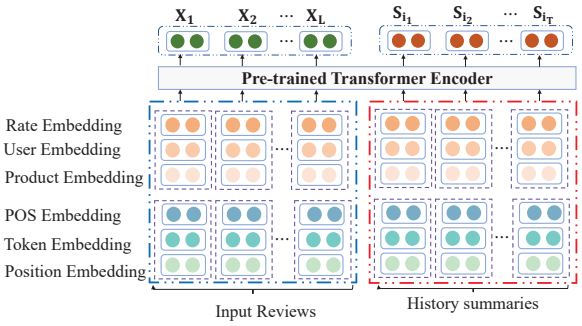

Figure 2: The personalized encoder framework of our method.

As a result, our method could select the salient content of the input review, which is not only based on semantic information but also reflect user and product characteristics After the personalized encoder, we could obtain the input review representations $\mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_L\}$ and history summaries representations $\mathbf{S} = \{\mathbf{S}_1, \cdots, \mathbf{S}_M\}$, where $\mathbf{S}_i = \{\mathbf{S}_{i_1}, \cdots, \mathbf{S}_{i_T}\}$ is the representation of i-th history summary.

2.3 Interactive Information Selection

Based on the learned representations of the input reviews and history summaries, in this section, we propose an interactive information selection module to interactively model the input reviews and history summaries. As shown in Figure 3, this module intents to further identify the salient content of the input review in terms of history summaries, meanwhile selecting the important information of history summaries relevant to current summary generation.

Intuitively, some content of history summaries is less relevant to the main point of input review. For example, different users focus on different aspects of the current product, hence different history summaries of products have different relevance to the current summary generation. Therefore, we design a relevance attention mechanism that utilizes the input review as the query to select the relevant content from history summaries and it is calculated as follows:

$$Softmax(\frac{(\mathbf{X}\mathbf{W}_X^Q)^T(\mathbf{S}\mathbf{W}_S^K)}{\sqrt{d_e}})(\mathbf{S}\mathbf{W}_S^V), \quad (2)$$

where $\mathbf{W}_X^Q$, $\mathbf{W}_S^K$, $\mathbf{W}_S^V$ are learnable parameters. Then, history summaries having more semantic

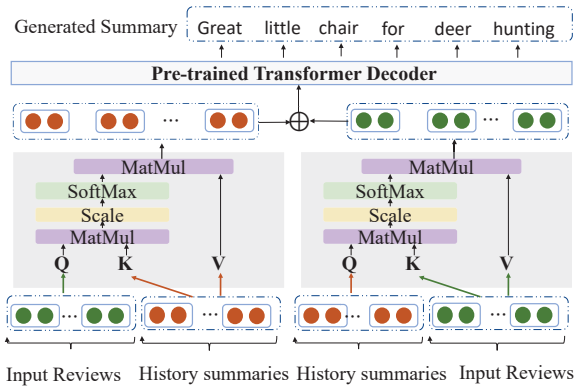

Figure 3: The decoder framework of our method. It first conducts interactive information selection for input review and history summaries, then generates personalized summaries based on the selected content.

similarity with input review would get more attention, which are then treated as an auxiliary feature to strengthen summary generation.

For input review, the personalized encoder has captured the internal salience of the input review by modeling the relatedness between words of the review via the self-attention mechanism. In fact, history summaries contain rich descriptions of the user and product characteristics, which conveys more semantic information than the user and product IDs. More specifically, history summaries of users reflect users' writing styles and purchasing preferences, and history summaries of products describe the main aspects that users are interested in. Therefore, we design another salience attention mechanism to capture the history summaries-aware salience for the input review and it is calculated as follows:

$$Softmax(\frac{(\mathbf{S}\mathbf{W}_S^Q)^T(\mathbf{X}\mathbf{W}_X^K)}{\sqrt{d_e}})(\mathbf{X}\mathbf{W}_X^V), \quad (3)$$

where $\mathbf{S}$ is the concatenation of all history summaries, $\mathbf{W}_S^Q$, $\mathbf{W}_X^K$, $\mathbf{W}_X^V$ are learnable parameters. In this way, our method could identify the salient content of the input review more effectively.

Finally, the concatenation of input review and history summaries is fed into the pre-trained transformer decoder which generates the target summaries $\hat{Y}$ word by word. The decoder also consists of stacked identical layers, in which there is an additional encoder-decoder self-attention to align the generation states and input sequences besides the two sub-layers in the encoder layer.

Dataset	Users	Products	Reviews
Movies and TV	123,960	50,052	1,697,471
Sports and Outdoors	35,598	18,357	296,214
Home and Kitchen	66,212	27,991	550,461

Table 1: Dataset statistics.

2.4 Model Training

For the summary generation task, we use the negative log-likelihood as the loss function (NLLLoss) to train the model:

$$\mathcal{L}_\phi(\hat{Y}|X,A) = \sum_{t=0}^{\hat{T}} -logP(\hat{y}_t)\ , \quad (4)$$

where $\hat{T}$ is the length of the generated review summary, $P(\hat{y}_t)$ is the probability distribution of the t-th word, and ϕ is model parameters.

3 Datasets and Experimental Settings

3.1 Datasets

In this section, we introduce the dataset statistics and hyperparameters settings in experiments. To validate the effectiveness of our method, we conduct extensive experiments on three real-world datasets from Amazon [1]:**Movies and TV**, **Sports and Outdoors**, and **Home and Kitchen**. Each sample of the dataset contains the user ID, product ID, rating, review, and summary text. Following previous work (Ma et al., 2018), we randomly select 1000 samples as testing and validation set separately and treat other samples in the dataset as the training dataset. In this paper, we only reserve the reviews given by active users to popular products, where each user and each product has at least K history reviews, where $K = 5$ for the *Sports* dataset, $K = 10$ for the *Home* dataset, and $K = 20$ for the *Movie* dataset. In the experiment, we utilize $M = 20$ history summaries. For reviews that have more than M history summaries, we select top-M history summaries that have more common words with the input review. The maximum length of reviews and summaries are set to $L = 200$ and $T = 15$ respectively. The dataset statistics are listed in Table 1.

3.2 Baselines

In this section, we compare our method with several state-of-the-art review summarization methods. **(1)** the methods without user and product

[1] http://jmcauley.ucsd.edu/data/amazon/

Dataset	Movie			Sports			Home		
Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Transformer	10.62	2.14	10.30	14.55	4.05	14.43	13.56	2.24	13.29
S2S+Attn	11.58	2.93	11.33	15.30	4.64	15.15	13.95	4.13	13.78
PGN	12.63	3.86	12.25	16.36	5.38	16.19	15.30	4.50	15.16
T5	8.14	1.91	7.58	8.91	2.53	8.33	9.49	2.59	8.8
T5-FT	13.32	5.94	12.84	18.08	7.41	17.82	16.3	6.74	16.12
HSSC	12.35	3.55	12.09	15.49	4.11	15.28	14.39	4.28	14.07
Dual-view	13.13	3.84	12.79	16.68	5.15	16.35	15.49	5.11	15.23
USN	13.58	4.08	13.20	14.98	5.10	14.85	13.37	2.94	13.23
memAttr	13.73	4.29	13.34	18.54	7.25	18.33	17.38	5.78	17.18
TRNS	15.38	4.96	14.96	19.78	6.12	19.54	18.03	5.72	17.88
PPRS	**16.49**	**7.18**	**16.08**	**19.92**	**8.07**	**19.63**	**18.52**	**7.7**	**18.08**

Table 2: ROUGE performance on three datasets. The improvements of our proposed method over all baselines are significant with p-value< 0.05.

information: *Transformer* (Vaswani et al., 2017), *S2S-att* (Bahdanau et al., 2015), *PGN* (See et al., 2017). **(2)** the personalized review summarization methods with the user and product information: *HSSC* (Ma et al., 2018) and *Dual-view* (Chan et al., 2020) jointly optimize review summarization and sentiment classification by taking rating as sentiment label; *USN* (Li et al., 2019a), *memAttr* (Liu and Wan, 2019) and *TRNS* (Xu et al., 2021) jointly consider discrete attributes and history text. **(3)** the methods based on the pre-trained language model: we compare our method with the original *T5* (Raffel et al., 2020) method and *T5-FT* which fine-tunes T5 on the recommendation datasets by generating summaries from product reviews.

3.3 Implementation Details

The hyper-parameters in our model are tuned from the validation dataset. The dimension of the hidden state d_e and attribute embedding (e.g., user ID) size d_a is set to 512. We use t5-small [2] to initialize the encoder and decoder parameters. We utilize the AdamW (Loshchilov and Hutter, 2019) algorithm to optimize our model and the learning rate is 0.0004. For the parameters in the training, we set the batch size to 32. For baselines, we use open source code for *HSSC*, *Dual-view*, *memAttr*, and *Transformer*. And we implement *S2S-att*, *PGN*, *USN*, *TRNS* and keep the same setting as the original papers. As for metrics, we use the widely used metrics ROUGE (Lin, 2004) [3] to evaluate the performance of our model on summary generation, including ROUGE-1, ROUGE-2, and ROUGE-L. Finally, the experiment platform is GeForce GTX 1080Ti with 128GB memory; we independently repeat each experiment 5 times and present the average performance.

[2]https://huggingface.co/t5-small

[3]https://github.com/chakki-works/sumeval

4 Experiments

4.1 Performance Evaluation

The results are listed in Table 2, from which we could have the following observations.

First, our method outperforms methods without the user and product information (e.g., *Transformer*) by a large margin. The main reason is that user and product features are crucial to generate high-quality summaries for product reviews in the recommendation scenario. During these methods, PLMs-based methods (i.e., *T5* and *T5-FT*) achieve better performance than other models. This is because these methods have strong text understanding ability obtained from the pre-training process which is helpful to identify the salient contents of the input review more accurately.

Second, our method achieves better performance than other methods that also leverage personalized information of users and products. Because our method could conduct more effective salience estimation by jointly considering semantic information and personalized information in the encoder and information selection, which further boosts summary generation. It should be noted that methods fusing history texts and discrete attributes (e.g., *TRNS*, *memAttr*) perform better than methods only based on discrete (e.g., *HSSC*, *Dual-view*). The possible reason is history texts convey more semantic information about user writing style and product characteristics which are clues to generate personalized summaries.

Third, we can see that our method performs better than PLMs-based baselines (i.e., *T5* and *T5-FT*). In fact, *T5* ignores the domain knowledge in recommendations resulting in poor performance, while *T5-FT* achieves better performance by learn-

ing the domain knowledge by fine-tuning PLMs on the reviews dataset. However, *T5-FT* still performs poorly than our method since it ignores the user preferences and product characteristics which play a crucial role in review summarization. Our method could incorporate this information into the salience calculation of the input reviews effectively and feed the relevant history summaries into the decoder, which both contribute to the improvement of our method. These results indicate that our method could fine-tune the pre-trained language model more effectively.

Models	ROUGE-1	ROUGE-2	ROUGE-L
PPRS	**19.92**	**8.07**	**19.63**
PPRS w/o H	19.82	7.93	19.54
PPRS w/o R	19.77	7.64	19.46
PPRS w/o S	19.57	7.33	19.31
PPRS w/o E	19.24	7.80	18.92
T5-FT	18.08	7.41	17.82

Table 3: Ablation experiments on the Sports dataset.

4.2 Ablation Study

To verify the effectiveness of important components of our method, in this section, we conduct an ablation study experiment by removing each component separately. Specifically, (1): **"w/o H"** denotes removing the history summaries information in the decoder module; (2): **"w/o H"** denotes removing the salience attention in the interactive information selection module; (3): **"w/o S"** denotes removing the interactive information selection module; (4): **"w/o E"** denotes removing the newly denoted embeddings (i.e., rating, user ID, product ID, and part-of-speech) in the personalized encoder module. The results are shown in Table 3 and we have the following observations.

We can see that removing any component would make the performance decline. First, removing discrete attributes and linguistic features makes our method could not identify the salient content of the input reviews that reflect user preferences and product characteristics effectively, resulting in the loss of personalized information in the generation process and achieving worse performance. Second, removing the information selection module and history summaries makes our method could not identify the salient content of input reviews more accurately and lost the important information from the history summaries, hurting the performance on summary generation. In addition, all variants outperform *T5-FT* which directly fine-tunes *T5* on the review dataset without considering the personalized features. In all, these results validate the effectiveness of these important components.

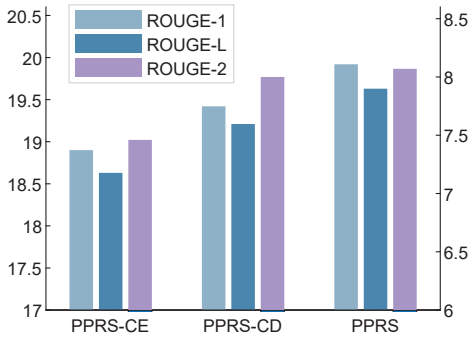

Figure 4: Performance of different mechanisms to utilize history summaries to expand the capacity of models.

4.3 Discussion

In this section, we conduct experiments to analyze the influence of different strategies which incorporate user preferences and product characteristics into the summary generation process.

Firstly, we design two variants to explore the effectiveness of different mechanisms to utilize the history summaries to expand the capacity of models. (1): **"PPRS-CD"** encodes the input review and history summaries separately, then feeds the concatenation of learned semantic representations into the decoder. (2):**"PPRS-CE"** directly feed the concatenation of the input review and history summaries into the encoder-decoder framework to produce summaries. The results are listed in Figure 4.

We can see that "PPRS-CD" performs worse than "PPRS" after replacing the interactive information selection with a concatenation operation. Because there is generally some irrelevant content in history summaries which might make the decoder confused about the input text and hurt the quality of the generated summaries. Then, "PPRS-CE" also achieves worse performance than "PPRS" after conducting information fusion in the encoder module. The possible reason is encoder could not distinguish the input reviews and history summaries and further fails to learn accurate semantic representations for them respectively, resulting in the performance declines.

Secondly, we design two variants to explore different strategies to utilize the discrete personalized information,i.e., rating, user and product IDs. (1): **"PPRS-AG"** utilizes rating, user and product id

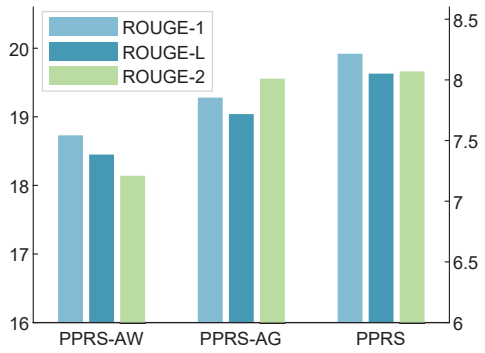

Figure 5: Performance on different mechanisms to leverage the discrete attributes (i.e., IDs and rating).

embeddings as a gate to select the relevant information from the text (i.e., input review and history summaries) representations after the encoder. (2): **"PPRS-AW"** treats rating, user and product id as special words and adds these embeddings to the beginning of the reviews and history summaries. The results are listed in Figure 5.

We can observe that the performance of summary generation begins to decline when applied the gate mechanism to conduct information selection in "PPRS-AG". This is because in this case, "PPRS-AG" could not conduct salience estimation well without the deep interaction between the user id and input text, leading the decoder could not generate more accurate summaries. Besides, "PPRS-AW" performs worst compared with other methods. The main reason is it could not identify the salient words in terms of user preferences and product characteristics in the encoder module effectively, which makes 'PPRS-AW" could not learn more comprehensive representations of input review and history summaries. However, our method aligns each word with these discrete attributes in the embedding layer which could incorporate the essential characteristics of users and products into the text encoding process effectively and further boosts the generation process.

4.4 Human Evaluation

In this section, we perform a human evaluation to further evaluate the performance of our model. Specifically, we define three metrics: **(1) Informativeness** evaluates whether the generated summaries convert the main content of the input review. **(2) Accuracy** evaluates whether the generated summaries are consistent with the sentiment tendency reflected in the input review. **(3) Readability** evaluates whether the generated summaries are grammatically correct and easy to understand. It is difficult to develop automatic evaluation methods for these metrics. Hence, we randomly sample 100 cases and invite 5 human volunteers to read and rate all generated summaries, where 1 means "very bad" and 5 means "very good". The scores are averaged across all volunteers and cases. The results are listed in Table 4 and we have the following observations.

Methods	Informativeness	Accuracy	Readability
PGN	3.75	3.67	4.05
TRNS	4.27	4.10	4.16
T5-FT	3.96	3.98	4.35
PPRS	**4.33**	**4.32**	**4.48**

Table 4: Human evaluation on *Sports* dataset.

First, our method outperforms other methods on *Informativeness* and *Accuracy*. Because our method could incorporate user preferences and product characteristics into the salience calculation of input reviews more effectively, which is helpful to capture the main content and keep the sentiment consistent with the input review. Second, the generated summaries of our method are more readable than others (e.g., *T5-FT*). The main reason is our method not only has rich grammar knowledge and text generation ability obtained from the pre-trained process, but also jointly considers user writing style by taking history summaries as auxiliary features. In summary, these results show that our method could generate high-quality summaries.

4.5 Case Study

In this section, we conduct a case study and we list several generated summaries of our method in Table 6. And we have the following observations.

First, generated summaries preserve the main contents of the input review and they have the same sentiment tendency. For example, in the first case, generated summaries and reviews both mention "training ammo" and convey a positive opinion (i.e., "Recommended") towards the product. Second, generated summaries are semantically similar to the corresponding reference summaries, such as, they both mention that "Doesn't work for recoil" in the second case. Third, generated summaries reflect user preferences and product characteristics. For example, in the third case, the generated summary contains two main features of the product (i.e., "little" and "great hunting") and indicates that

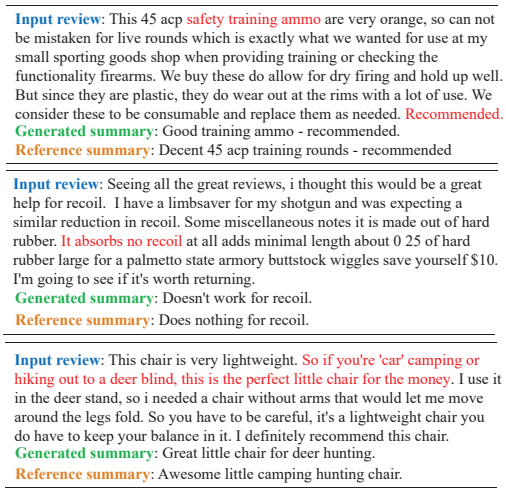

Figure 6: Examples of generated and reference reviews.

the corresponding user cares more about the "quality". In all, our method could generate coherent and personalized summaries for product reviews.

5 Related Work

5.1 Personalized Review Summarization

Personalized review summarization is an important task in the recommender system, which aims to generate brief summaries for product reviews. Different from the previous text summarization (Gehrmann et al., 2018; Li et al., 2020; Zhang et al., 2019), product reviews usually have various personalized information (e.g., rating, user and product IDs, and history text, etc.) which plays a crucial role in summary generation (Yang et al., 2018; Dong et al., 2017).

Recently, some approaches (Ganesan et al., 2010; Xiong and Litman, 2014; Carenini et al., 2013; Di Fabbrizio et al., 2014; Liu and Wan, 2019; Li et al., 2019b,c; Chan et al., 2020) are proposed for review summarization. Some methods incorporate the discrete attributes (e.g., rating, user and product IDS) into salient information selection. Li et al. (2019a) design a selective mechanism that utilizes user embedding to select user-preference words and generate a personalized summary by incorporating user-specific vocabulary. In addition, some methods also leverage aspect information to enhance review summarization (Yang et al., 2018; Tian et al., 2019). However, most of them ignore the joint consideration of the discrete attributes and history text. Therefore, Liu et al. (2019) calculate the semantic similarity between input review and history review to aggregate history summaries into context vectors which are then utilized to generate summaries. Xu et al. (Xu et al., 2021) conduct deep interaction between input reviews and history summaries to infer the important parts among history summaries and generate personalized summaries by reasoning over the user-specific memory.

5.2 Pre-trained Language Model

Recently, pre-trained language models (PLMs) have advanced the performance of various NLP tasks, such as sentiment analysis (Yu et al., 2021; Wu and Shi, 2022), text summarization (Liu and Lapata, 2019; Xiao et al., 2020; Oved and Levy, 2021), etc. Liu and Lapata (2019) propose to conduct summary generation in both extractive and abstractive modeling paradigms by utilizing BERT (Kenton and Toutanova, 2019) as an encoder to learn text representations. Oved and Levy (2021) generate opinion summaries for products by aggregating a set of reviews for the given product and significantly reduce the self-inconsistencies between multiple history reviews. However, these methods might perform poorly in personalized review summarization, since they ignore the rich characteristics of users and products which is important to generate high-quality summaries for reviews. In this paper, we propose to fine-tune PLMs to conduct more effective salience estimation for input reviews by jointly considering semantic information and personalized features of users and products.

6 Conclusion

In this paper, we propose a novel review summarization method based on the pre-trained language models. The core of our method is fine-tuning the pre-trained language models by considering the user preferences and product characteristics. Especially, we design a personalized encoder to learn representations for the input reviews and each history summary separately by incorporating the user and product characteristics into the encoder module. Additionally, we propose an interactive information selection module to further identify the salient content of the input review and select the relevant information from history summaries. Experimental results show that our method achieves better performance than competitive baselines.

7 Limitations

Our method has some limitations that we would like to explore in the future. Firstly, our method

is based on the PLMs which require large GPU resources to train and infer models. We would like to adopt knowledge distillation technology to reduce the number of model parameters while keeping the performance as much as possible. Secondly, the summary generation process still lacks enough controllability even though we incorporate various features of users and products into the saliency estimation and auxiliary inputs of the decoder. In the future, we explore aggregating the characteristics of users and products into the decoder layers to make the generation process more controllable.

Acknowledgement

This work was supported by the Shenzhen Sustainable Development Project under Grant (KCXFZ20201221173013036).

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6578–6593. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. 2013. Multi-document summarization of evaluative text. *Computational Intelligence*, 29(4):545–576.

Hou Pong Chan, Wang Chen, and Irwin King. 2020. A unified dual-view model for review summarization and sentiment classification with inconsistency loss. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1191–1200. Association for Computing Machinery.

Giuseppe Di Fabbrizio, Amanda Stent, and Robert Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 54–63.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632.

Kavita Ganesan, Cheng Xiang Zhai, and Jiawei Han. 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In *23rd International Conference on Computational Linguistics, Coling 2010*.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1602–1613.

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

Junjie Li, Haoran Li, and Chengqing Zong. 2019a. Towards personalized review summarization via user-aware sequence network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6690–6697.

Junjie Li, Xuepeng Wang, Dawei Yin, and Chengqing Zong. 2019b. Attribute-aware sequence network for review summarization. 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 2991–3001.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. A survey of pretrained language models based text generation. *CoRR*, abs/2201.05273.

Piji Li, Lidong Bing, Zhongyu Wei, and Wai Lam. 2020. Salience estimation with multi-attention learning for abstractive text summarization. *arXiv preprint arXiv:2004.03589*.

Piji Li, Zihao Wang, Lidong Bing, and Wai Lam. 2019c. Persona-aware tips generation. In *The World Wide Web Conference*, pages 1006–1016. ACM.

Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *SIGIR*, pages 345–354. ACM.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hui Liu and Xiaojun Wan. 2019. Neural review summarization leveraging user and product information. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2389–2392.

Yang Liu and Mirella Lapata. 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 (EMNLP-IJCNLP)*, pages 3730–3740.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4251–4257.

Nadav Oved and Ran Levy. 2021. Pass: Perturb-and-select summarizer for product reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 351–365.

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

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Yufei Tian, Jianfei Yu, and Jing Jiang. 2019. Aspect and opinion aware abstractive review summarization with reinforced hard typed decoder. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2061–2064.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hui Wu and Xiaodong Shi. 2022. Adversarial soft prompt tuning for cross-domain sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2438–2447.

Liqiang Xiao, Lu Wang, Hao He, and Yaohui Jin. 2020. Modeling content importance for summarization with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3606–3611.

Wenting Xiong and Diane Litman. 2014. Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1985–1995.

Hongyan Xu, Hongtao Liu, Pengfei Jiao, and Wenjun Wang. 2021. Transformer reasoning network for personalized review summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1452–1461.

Min Yang, Qiang Qu, Ying Shen, Qiao Liu, Wei Zhao, and Jia Zhu. 2018. Aspect and sentiment aware abstractive review summarization. In *Proceedings of the 27th international conference on computational linguistics*, pages 1110–1120.

Jianfei Yu, Chenggong Gong, and Rui Xia. 2021. Cross-domain review generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4767–4777.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

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

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

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3,4

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
The human evaluation for review summarization is simple to conduct, including Informativeness, Accuracy, and Readability. Hence, we do not design text instruction specifically.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
5

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

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