# Probing Product Description Generation via Posterior Distillation

**Haolan Zhan,[1]* Hainan Zhang,[2†] Hongshen Chen,[2†] Lei Shen,[3,4] Zhuoye Ding,[2] Yongjun Bao,[2] Weipeng Yan,[2] Yanyan Lan[3,4†]**

[1]Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]Data Science Lab, JD.com, Beijing, China
[3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[4]University of Chiense Academy of Sciences, Beijing, China
zhanhaolan316@gmail.com, zhanghainan6@jd.com, ac@chenhongshen.com, {shenlei17z, lanyanyan}@ict.ac.cn

## Abstract

In product description generation (PDG), the user-cared aspect is critical for the recommendation system, which can not only improve user's experiences but also obtain more clicks. High-quality customer reviews can be considered as an ideal source to mine user-cared aspects. However, in reality, a large number of new products (known as long-tailed commodities) cannot gather sufficient amount of customer reviews, which brings a big challenge in the product description generation task. Existing works tend to generate the product description solely based on item information, i.e., product attributes or title words, which leads to tedious contents and cannot attract customers effectively. To tackle this problem, we propose an adaptive posterior network based on Transformer architecture that can utilize user-cared information from customer reviews. Specifically, we first extend the self-attentive Transformer encoder to encode product titles and attributes. Then, we apply an adaptive posterior distillation module to utilize useful review information, which integrates user-cared aspects to the generation process. Finally, we apply a Transformer-based decoding phase with copy mechanism to automatically generate the product description. Besides, we also collect a large-scare Chinese product description dataset to support our work and further research in this field. Experimental results show that our model is superior to traditional generative models in both automatic indicators and human evaluation.


Figure 1: A product description example from our dataset.

## Introduction

In E-commerce, the goal of online product recommendation system is to post suitable commodities to customers and stimulate their purchasing behaviors. However, ranking the products and display them to users can no longer meet the requirements of customers (Zhang et al. 2019; Gong et al. 2019; Chen et al. 2019a). While browsing the recommendation system, customers face the problem of information explosion. To save costs and find products in need straightforwardly, customers would like to see some refined product descriptions rather than complex product details, as shown in Figure 1. Therefore, it is important to present product characteristics in a product description for the E-commerce system to help customers learn the recommended products directly.

Furthermore, products with well-written description is capable to attract more customer's attentions. For instance, as showed in Figure 1, the product description below the product depicts some user-cared aspects, i.e., "*comfortable and soft*" and "*improving the wearing experience*", which can arouse customer's interests and encourage them to buy it. With this appealing user-cared description, customers could select their interested products easily and feel more satisfied with the entire recommending process. In a word, generating an user-cared product description is an important and practical research problem in E-commerce scenario.

High-quality customer reviews are an ideal source to mine user-cared aspects (Pecar 2018). The customer post their re-

*Work done at Data Science Lab, JD.com.
†Corresponding authors.

views of a product, which naturally shows their most cared aspects. However, in reality, lots of new products (long-tailed products) cannot gather sufficient amount of customer reviews. We make some statistics of customer reviews in Table 1. In category *Shoes&Clothes*, there are more than 66.3% commodities have less than 10 reviews and the average number of reviews is only 18.4. The data shows that long-tailed phenomenon of customer reviews is obvious in the E-commerce system. That is to say, a large number of products lack enough corresponding reviews but we still need to generate product descriptions for them.

Recently, most existing methods (Zhang et al. 2019; Li et al. 2020) consider item contents, such as product image, text, attributes and title, as their source to generate the product description for long-tailed products. Obviously, the generated descriptions may be tedious and cannot attract customers effectively since they ignore user's experience. To enhance the effectiveness of user-cared aspects, other researchers (Chen et al. 2019a,b) propose to incorporate customer's personalized profiles and/or external product knowledge from Wikipedia to generate product descriptions. However, the personalized data is too sparse and thus hard to represent and utilize. On the other hand, these methods also cannot deal with the long-tailed products because the personalized data is inaccessible.

To tackle this problem, we propose an **A**daptive **P**osterior **D**istillation model based on **T**ransformer architecture (APDT), which can utilize user-cared aspects from customer reviews, and then incorporate these aspects into the generation process of product descriptions. Specifically, we first extend the self-attentive Transformer encoder to encode product items (title and attributes) and reviews. Then, we apply an adaptive posterior distillation layer to utilize effective review information. In this layer, product title and attributes representation are fused into item representation through feature fusion module at first. Then, the review representation is updated by interacted with item representation. During training phase, item and review representations are sent into decoder layer separately. KL divergence loss is employed in the distillation process to approximate item and review representations. Finally, we apply a Transformer decoding phase with copy mechanism to automatically generate product descriptions. Besides, to enhance the coherence between generated description and ground truth, we also employ a coherence-enhanced function during training.

In our experiments, to evaluate our automatic product description generation task, we construct a new Chinese dataset from JD.com, one of the biggest e-commerce platform in China. This dataset contains 345,799 pairs of item content and description. The results on this dataset show that our model outperforms the state-of-the-art generative baselines, in terms of both automatic and human evaluations.

Our contributions are listed below: 1) We propose an adaptive posterior distillation Transformer model to tackle the long-tailed commodities problem in product description generation task. 2) We collect a large-scale Chinese product description dataset for this research point. 3) Experimental results on this dataset validate the effectiveness of our proposed model.

| Category | #Products | #Review(Avg) | #<10 |
|---|---|---|---|
| Shoes&Clothes | 143,941 | 18.4 | 66.3% |
| Digital | 108,236 | 15.7 | 58.7% |
| Homing | 93,622 | 21.6 | 68.2% |

Table 1: Statistics of customer review information.

# Related Work

## Text Generation in E-commerce

Text generation in E-commerce aims at improving customer's online shopping experience. Several novel and challenging tasks are proposed, including short title generation (Zhang et al. 2019), product description generation (Chen et al. 2019a) and recommendation reason generation (Zhan et al. 2020). The motivation of STG is to concisely display short product titles on limited screen of mobile phones. Gong et al. (2019) firstly proposed the short title generation task for e-commerce, which automatically generates short title by directly extracting essential information from original long title. Wang et al. (2018) proposed a multi-task learning approach by using external searching log data as additional task to facilitate key words extraction process. Furthermore, Zhang et al. (2019) considered a multi-source approach incorporating multi-modal information with generative adversarial networks. As for product description generation task, early work focuses on template-based generation approaches that incorporates statistical methods (Wang et al. 2017). With the evolution of neural network methods, RNN and Transformer are applied in this task. Chen et al. (2019a) proposed a personalized knowledge transformer model to generate the product description. Their methods utilized the item-based features, i.e., product image, attributes and title, and external knowledge base, such as Wikipedia. However, the external knowledge base risks introducing noise, which may hurt the effectiveness of generating personalized product description.

## Personalized Content Generation

Personalized content generation has attracted research interest in various domains, e.g., the automatic generation of marketing messages (Roy et al. 2015; Chen et al. 2020), persuasive message (Ding and Pan 2016), poetry generation (Shen, Guo, and Chen 2020), argument generation (Carenini and Moore 2006) and dialogue generation (Shen and Feng 2020; Feng et al. 2020a; Shen, Feng, and Zhan 2019; Shen et al. 2021). With the support of user preferences, the effectiveness has increases. Recently, Krishna et al. (2018) presented a framework for the summary generation that takes into consideration the linguistic preferences of the specific audience. Reichelt et al. (2014) showed that personalized information of learning materials can increase motivation and learning outcomes. Zander et al. (2015) studied the effect of personalization on students' attention allocation using some eye-tracking methods, and find that the personalized parts of reading materials are more attractive. In the field of E-commerce, Elad et al. (2019) proposed an extractive method to select sentences and then generate personalized product

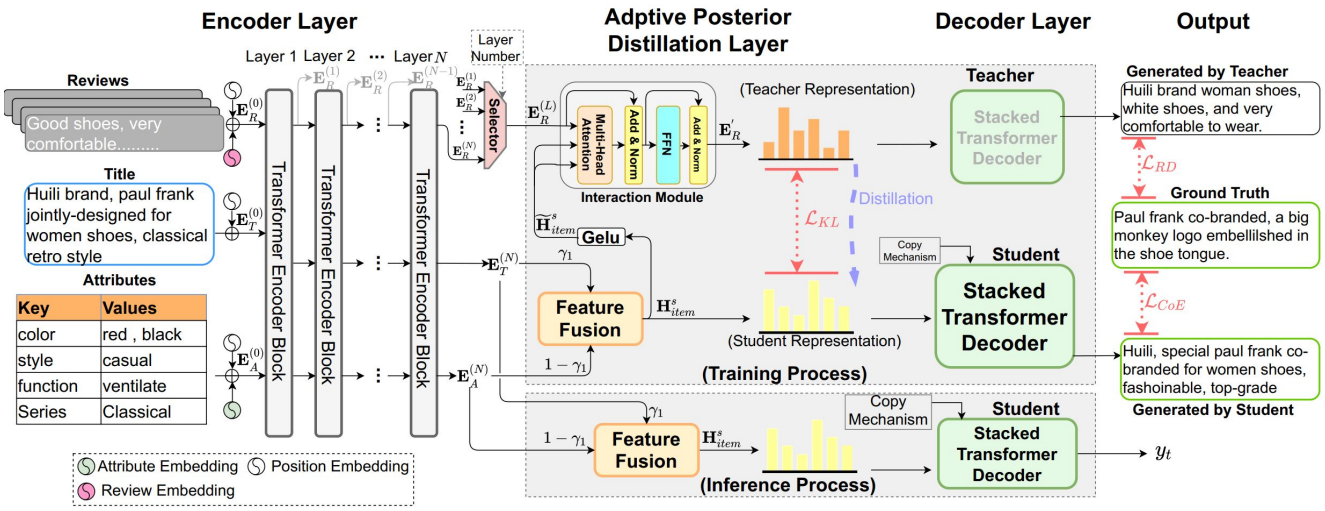



Figure 2: The architecture of APDT model. Here we omit the architecture inside the "Transformer Encoder Block" and "Stacked Transformer Decoder", and refer the readers to Vaswani et al. (2017) for more details.

description. Chen et al. (2019b) built a bridge between personalized outfit generation and recommendation by considering both user preferences and individual items. To the best of our knowledge, our method takes the first attempt to introduce user-cared aspects for product generation task.

## Proposed Method

### Problem Formulation

Given a product title, attribute sets and its corresponding customer reviews sets, the product description generation (PRG) task aims at utilizing inherent product information at first, and then identifying the privilege of customer's preference. Finally, coherent and appealing product descriptions will be generated.

Formally, given a product title $T$ composed of a sequence of words $\{t_1, ..., t_N\}$, a group of attributes $A$ contains $M$ pairs of key-values, i.e., $A = \{\{k_1, a_1\}, ..., \{k_M, a_M\}\}$ and a relative customer review set defined as $\{R_1, ..., R_n\}$, the PDG task attempts to learn a generative model $G(\cdot)$. Each review sentence in $R$ is defined as $R_n = \{r_n^1, \cdots, r_n^L\}$, where $r_n^i$ is the $i$-th word in the sentence of $R_n$ and $L$ is the max length for review sentence. The corresponding generated product description is defined as $Y = \{y^1, \cdots, y^S\}$, where $y^i$ is the $i$-th word and $S$ is the max length of product description. With sequence-to-sequence framework, this can be formulated as follows:

$$P(Y|T, A, R; \theta) = \prod_{t=1}^{|S|} P(y_t|y_{<t}, T, A, \{R_i\}_{i=1}^n; \theta),$$

where $\theta$ is the parameter, and $y_{<t}$ denotes the previously generated words.

### Overview of Our Model

As shown in Figure. 2, our model consists of three main layers: (1) an encoding layer, (2) an adaptive posterior distillation layer, and (3) a decoding layer with copy mechanism. The encoding layer employs a stacked Transformer encoder module (Vaswani et al. 2017) to encode context including attributes, titles and customer reviews. The adaptive posterior distillation layer contains feature fusion module, interaction module, and the teacher representation and student representation learning module. The decoding layer contains a stacked Transformer decoder to generate response $y_t$ token by token.

### Encoding Layer

In the encoding layer, we apply the Transformer encoder module with different position encoding mechanism to the title, attributes and reviews, separately.

Given a product title $T = \{t_1, ..., t_N\}$ as the input, the initial word embedding and position embedding vectors are represented as $WE(T)$ and $PE(T)$ respectively. The initial input title representations $\mathbf{E}_T^{(0)}$ is the sum of word and position embedding at the first layer:

$$\mathbf{E}_T^{(0)} = WE(T) + PE(T).$$

At the $l$-th layer, the output representation is defined as below:

$$\mathbf{E}_T^{(l)} = \text{FFN}(\text{MHA}(\mathbf{E}_T^{(l-1)}, \mathbf{E}_T^{(l-1)}, \mathbf{E}_T^{(l-1)})),$$

where $\mathbf{E}_T^{(l)}$ denotes the output representations after the $l$-th layer. The sub-layer $\text{FFN}(\cdot)$ is a position-wise fully connected feed-forward network, and $\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ (Vaswani et al. 2017) is a multi-head attention function. We refer the readers to Vaswani et al. (2017) for more details.

For attributes context, we apply a unique attribute embeddings (AE), to adapt to its structured data format. Attribute embeddings are employed to differentiate the key-value pairs in the attribute sets. Therefore, inside the key-value pair, the words share same attribute embeddings. The

initial representation of attributes representation $\mathbf{E}_A^{(0)}$ and encoding phases are defined as:

$$\mathbf{E}_A^{(0)} = WE(A) + PE(A) + AE(A),$$
$$\mathbf{E}_A^{(l)} = \text{FFN}(\text{MHA}(\mathbf{E}_A^{(l-1)}, \mathbf{E}_A^{(l-1)}, \mathbf{E}_A^{(l-1)})).$$

For customer review representations, given a review set $\{R_1, ..., R_N\}$ as the input, we firstly concatenation all the words as a sequence. Then, we will apply a review embeddings (RE) to differentiate the review sentences. The initial representation of review sentences $\mathbf{E}_R^{(0)}$ and encoding phases are defined as:

$$\mathbf{E}_R^{(0)} = WE(R) + PE(R) + RE(R),$$
$$\mathbf{E}_R^{(l)} = \text{FFN}(\text{MHA}(\mathbf{E}_R^{(l-1)}, \mathbf{E}_R^{(l-1)}, \mathbf{E}_R^{(l-1)})).$$

## Adaptive Posterior Distillation Layer

Inspired by the knowledge distillation's success (Hinton, Vinyals, and Dean 2015; Feng et al. 2020b) on model compression and knowledge transfer, we propose an adaptive posterior distillation layer to transfer user-cared aspects in review information (*teacher*) to item representation (*student*), which contains title and attributes information.

During the posterior training process, we design an individual training objective for reviews (*teacher*) information, in order to enhance the semantic coherence between review information and target product description.

**Student representation**: Firstly, we define an item representation to combine title and attributes representation. The item representation (*student*) is defined as:

$$\mathbf{H}_{item}^s = \gamma_1 \mathbf{E}_T^{(N)} + (1 - \gamma_1)\mathbf{E}_A^{(N)},$$

where $\gamma_1 \in [0, 1]$ is a parameter, and $\mathbf{E}_T^{(N)}$ and $\mathbf{E}_A^{(N)}$ is the final representation of title and attributes that output from the $N$-th encoder layer, also known as the last encoder layer.

**Teacher representation**: Given the reivew representation $\mathbf{E}_R^{(N)}$, we firstly apply an interaction module to incorporate item information (title and attributes) into the representation of review information. The interaction module is designed to highlight user-cared aspects in reviews, with the assistance of item representation. It is a one layer mutli-head attention following with a feed-forward sub-layer.

To make all dimension of representation matrix compatible, we perform a non-linear projection of the parameters in student representation $\mathbf{H}_{item}^s$ before fed into interaction module. Therefore, the updated item representation (student) $\widetilde{\mathbf{H}}_{item}^s$ and updated review representations (*teacher*) $\mathbf{E}_R^{'}$ are defined as:

$$\widetilde{\mathbf{H}}_{item}^s = Gelu(W_1 \mathbf{H}_{item}^s + b_1),$$
$$\mathbf{E}_R^{'} = \text{FFN}(\text{MHA}(\widetilde{\mathbf{H}}_{item}^s, \widetilde{\mathbf{H}}_{item}^s, \mathbf{E}_R^{(L)})),$$

where $W_1$, $b_1$ are the parameters, and $Gelu$ (Gaussian Error Linear Unit) (Hendrycks and Gimpel 2016) is the non-linear projection function. $\mathbf{E}_R^{(L)}$ is the $L$-th encoder layer output, and $L \in [1, N]$. $L$ is a variable and set by human in experiments. As shown in Figure 2, Selector module (pink) is used to select the $L$-th encoder layer output.

we apply a stacked Transformer decoder layer to output $p_l^r(y_t)$ as the probability of token $y_t$ generated by the teacher model at the $t$-th step. Therefore, the review distillation training objective $\mathcal{L}_{RD}(\theta)$ as:

$$\mathcal{L}_{RD}(\theta) = -\sum_{t=1}^{S} \log p_l^r(y_t|\mathbf{E}_R^{'}; \theta)$$

where $p_l^r(y_t)$ is the representations generated by the $\mathbf{E}_R^{'}$ respectively for a token $y_t$.

During the posterior training phase, in order to approximate the distributions of student and teacher representation, we introduce the KL divergence loss, to measure the proximity between the prior item (student) representation and the posterior review (student) representation. The KL-divergence is defined as follows:

$$\mathcal{L}_{KL}(\theta) = D_{KL}(p^s(y_t|\mathbf{H}_{item}^s)||p_l^r(y_t|\mathbf{E}_R^{'}); \theta),$$

where $\theta$ denotes the model parameters.

In the inference process, we only keep the well-trained prior module, and then feed the item representation $\mathbf{H}_{item}^s$ into decoder layer.

## Decoding Layer

In the decoding layer, we apply a stacked Transformer decoder module equipped with a copying mechanism (See, Liu, and Manning 2017) to generate product description. We feed the product representation $\mathbf{H}_{item}^s$ into decoder layer. Specifically, the probability of generating token $y_t$ at $t$-th step is modeled as:

$$P(y_t) = \lambda_1 P_{vocab}(y_t|\mathbf{H}_{item}^s) + \lambda_2 P_{cp}(y_t|\mathbf{E}_T) + \lambda_3 P_{cp}(y_t|\mathbf{E}_A)$$

where $P_{cp}(y_t|\mathbf{E}_T)$ derives the copying probability from title words. The copy mechanism is defined as follows:

$$P_{cp}(y_t|\mathbf{E}_T) = \sum_{i:t_i=y_t} \alpha_{t,i},$$

and $P_{cp}(y_t|\mathbf{E}_A)$ derives the copying probability from attributes words, which is calculated in a similar way. $P_{vocab}(y_t|\mathbf{H}_{item}^s)$ is the output probability from a stack of Transformer decoder layers (Vaswani et al. 2017). $\lambda_1$, $\lambda_2$ and $\lambda_3$ are the coordination probability, which are estimated as follows:

$$[\lambda_1, \lambda_2, \lambda_3] = softmax(W_2 \mathbf{H}_{item}^s + W_3 \mathbf{E}_T + W_4 \mathbf{E}_A + b_2),$$

where $W_2, W_3, W_4, b_2$ are the parameters.

## Training Objectives

Besides applying the KL-divergence loss function for posterior distillation module, we also employ a **Co**herence-**E**nhanced Negative Log-Likelihood objective (CoE), which aims to help our model to generate words that seldom mentioned but coherent to user-cared aspects.

Different commodities which describe different aspects are always featured by the unique attribute values in the dataset. For example, a *clothes* category often has the attributes like 'texture', 'size'. The information in the unique

attributes is harder to capture than that in the common attributes like 'name', as the latter attributes are very frequent in the training set. We define the frequency of an attribute word $a_k$ as $f(a_k) = [freq(a_k)]^{-1}$ by calculating its frequency in the training set.

For a generated description $y^\star$, the coherence score between $y^\star$ and ground truth $y_g$ is calculated as follows:

$$Coh(y^\star) = \frac{\sum_{i=1}^{|y^\star|}(f(y_i^\star) \cdot \mathbb{B}\{y_i^\star \in y_g\})}{\sum_{i=1}^{|y|} f(y_i)},$$

where $f(\cdot)$ is the word frequency index, and $\mathbb{B}\{y_i^\star \in y_g\} = 1$ if word $y_i^\star$ is in the ground truth sentence $y$. If not, it equals to 0.

**Coherence-enhanced Function**: Different from previous models which only measures how well the generated sentences match the target sentences, we design a fused coherence-enhanced function $R_{fuse}$ which contains both the information coherence score and the ROUGE-L score (RG for short) of the generated descriptions.

$$R_{fuse}(y^\star) = \beta R_{Coh}(y^\star) + (1-\beta)R_{RG}(y^\star),$$

where $\beta$ is set to 0.4. $R_{Coh}(y^\star)$ is the coherence score between $y^\star$ and $y_g$, while $R_{RG}(y^\star)$ is designed to calculate the ROUGE score.

We apply a coherence-enhanced negative log-likelihood (CoE) as our training objective The training loss of the generation task is defined as:

$$\mathcal{L}_{CoE}(\theta) = -\sum_{t=1}^{|S|} R_{fuse}(y_{<t}) \cdot log(p(y_t|y_{<t}, T, A; \theta)),$$

Therefore, we optimize our all the following objectives jointly:

$$\mathcal{L}_{all}(\theta) = \alpha\mathcal{L}_{CoE}(\theta) + (1-\alpha)\Big(\mathcal{L}_{RD}(\theta) + \mathcal{L}_{KL}(\theta)\Big),$$

where $\alpha \in [0, 1]$, and it is used to weigh the contribution of different losses. A high value of $\alpha$ makes the student model focus more on generation task; whereas a relative lower value of $\alpha$ makes the student learn more from the teacher.

# Experiments

## Experimental Settings

**Dataset** We collect a large-scale Chinese product description generation dataset, named as JDPDG from JD.com[1], one of the biggest e-commerce platforms in China. Our dataset contains over 300 thousands product instances from the *Clothes&Shoes*, *Digital* and *Homing* categories. There are 104 kinds of products in *Clothes&Shoes* category,such as T-shirts and boots; 79 kinds of products in Digital, such as cameras and phones; 96 kinds of products in Homing, such as bowls and tobacco jars. Each commodity instance in our dataset includes a set of product information and a well-written product description. The set of product information contains a title, a group of attributes and a set of customer reviews. The product descriptions are written by thousands of qualified writers, with the reference of product title and attributes. The review information will be filtered at first, and only the the high-quality reviews are kept. The average number of words in each title, review and product description sentence are 13.8, 25.6 and 40.2, respectively. The average number of attribute keys in each product is 9.5, and for each key, its corresponding value contains 1 to 4 words. Table 2 shows more details about our dataset[2].

| Category | *Shoes&Clothes* | *Digital* | *Homing* |
|---|---|---|---|
| Training Pairs | 135,941 | 100,236 | 85,622 |
| Validation Pairs | 4000 | 4000 | 4000 |
| Test Pairs | 4000 | 4000 | 4000 |

Table 2: Data statistics for our proposed JDPDG dataset.

**Baseline Models** We compare our adaptive posterior distillation Transformer (APDT) model with several baseline models, including: (i) **PG-BiLSTM**: a bi-directional LSTM with pointer generator mechanism (See, Liu, and Manning 2017), (ii) **MS-Ptr**: a multi-source pointer network for short product title generation (Sun et al. 2018), (iii) **Transformer**: an encoder-decoder architecture relying solely on self-attention mechanisms (Vaswani et al. 2017), (iv) **HierTrans**: a hierarchical transformer for abstractive multi-document summarization tasks (Liu and Lapata 2019), (v) **EMA**: a unified text generation model for both structured and unstructured data with exponential moving average (EMA) technique (Shahidi, Li, and Lin 2020), (vi) **KOBE**: the state-of-the-art product description generation model with incorporated personalized knowledge attributes from external Wikipedia knowledge base (Chen et al. 2019a).

**Evaluation Metrics** We conduct both automatic and human evaluations. For automatic evaluation we follow previous PDG studies and use BLEU (Papineni et al. 2002) and ROUGE-L (Lin 2004). For human evaluation we randomly sample 200 examples from each test set. For each example, we ask six workers (both CS graduate students) to conduct a pairwise comparison between the product description generated by our APDT and other baselines. Specifically, each worker needs to give a preference in terms of three criteria: (1) Correctness, i.e., which description contains most correct information; (2) Diversity, i.e., which description looks more diversity; (3) Coherence, i.e., which description looks mostly coherent to the product. Each criterion is assessed with a score range from 1 (worst) to 4 (best).

**Implementation Details** We implement our model in OpenNMT[3] and train all models on the Tesla P40 GPUs with Pytorch (Paszke et al. 2019). For experimental models, the hidden units of all transformer-based models are set as 512 and the feed-forward hidden size is set as 1,024. The beam search size is set as 5 and length penalty as $\alpha$ = 0.4 (Wu

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

[2]https://github.com/jddsl/JDPDG

[3]https://github.com/OpenNMT/OpenNMT-py

| Model | *Clothes&Shoes* ROUGE-L | *Clothes&Shoes* BLEU | *Digital* ROUGE-L | *Digital* BLEU | *Homing* ROUGE-L | *Homing* BLEU |
|---|---|---|---|---|---|---|
| PG-BiLSTM (See, Liu, and Manning 2017) | 15.62 | 7.86 | 16.86 | 8.02 | 15.17 | 7.59 |
| MS-Ptr (Sun et al. 2018) | 15.95 | 7.98 | 16.54 | 7.79 | 15.72 | 7.74 |
| Transformer (Vaswani et al. 2017) | 16.38 | 7.83 | 16.64 | 7.63 | 16.58 | 7.31 |
| HierTrans (Liu and Lapata 2019) | 17.36 | 8.51 | 17.73 | 8.46 | 16.89 | 8.28 |
| EMA (Shahidi, Li, and Lin 2020) | 16.32 | 8.69 | 17.67 | 8.81 | 16.53 | 9.33 |
| KOBE (Chen et al. 2019a) | 19.07 | 9.27 | 18.97 | 9.32 | 18.72 | 9.41 |
| **APDT** (ours) | **20.41** | **10.36** | **19.95** | **10.08** | **19.68** | **10.13** |

Table 3: Automatic evaluation results on PDG dataset, including three different categories (%).

| Model | Correctness | Diversity | Coherence |
|---|---|---|---|
| PG-BiLSTM | 2.58 | **3.29** | 2.62 |
| MS-Ptr | 2.62 | 3.26 | 2.57 |
| Transformer | 2.19 | 2.87 | 2.30 |
| HierTrans | 2.45 | 3.04 | 2.54 |
| EMA | 2.63 | 2.91 | 2.75 |
| KOBE | 2.47 | 3.11 | 2.91 |
| **APDT** (ours) | **2.91** | 3.27 | **3.02** |

Table 4: Human evaluation on clothes&shoes category.

| Model | ROUGE-L | BLEU |
|---|---|---|
| **APDT** | **20.41** | **10.36** |
| – *Copy Mechanism* | 19.13 | 9.75 |
| – *Posterior Distillation* | 18.56 | 9.79 |
| – *Coherence Enhanced* | 19.87 | 10.02 |
| – *above all* | 17.95 | 9.13 |

Table 5: Ablation test on the clothes&shoes category (%).

et al. 2016). For LSTM-based models, the word dimension is set to 300 and the hidden nodes are set as 256 for the encoder and decoder. The dropout rate and smoothing factor are set as 0.1 (Fabbri et al. 2019). The initial learning rate is set to 0.001. The $\beta_1$ = 0.9 and $\beta_2$ = 0.998 are used for gradient optimization. We also apply warm-up trick over the first 8, 000 steps, and decay as in Vaswani et al. (2017). For hyper-parameters, we set $\gamma_1$, $\beta$ and $\alpha$ to 0.5, 0.4 and 0.5, respectively.

## Experimental Results

**Automatic Evaluation** The automatic evaluation results are shown in Table 3. Our proposed APDT model outperforms the best. Taking the ROUGE metrics as an example, the ROUGE-L value of the APDT in the *Clothes&Shoes* category is 20.41, which is significantly better than MS-Ptr, HierTrans, EMA and KOBE models i.e., 15.95, 17.36, 16.32 and 19.07. The BLEU metrics of our model is also higher than other baseline models,indicating that our model can generate more informative and fluency product description. We also conducted a significant test, showing that the improvement is significant, i.e., p-value $< 0.01$.

**Human Evaluation** We further conduct human evaluations to assess the proposed APDT model. Due to the limitation of pages, we only present the evaluation results on *clothes&shoes* category. But results on other two categories also show a similar trend. Table 4 summarizes the evaluation results. In the correctness criterion, our APDT model achieves a scores at 2.91, while other baseline models only get scores about 2.5. This result indicate that our model can generate more correct aspects. In the coherence criterion, APDT model can also achieves the best performance, indicating that APDT model can generate coherent and relevant information than baselines. We also employ Fleiss' kappa scores (Fleiss 1971) to measure the reliability between different annotators. The overall Fleiss' kappa score is 0.527.

## Case Study

To facilitate a better understanding of our model, we present some examples in Table 3. With the page limitation, we only present the generated production description from KOBE and our APDT model. For fair comparison, during inference process, we only send the product title and attributes sets into these two models. Review information are presented only for reference. As shown in Table 3, our proposed APDT model generates more aspects of product with considering customer review information. For example, the product description generated by KOBE model can cover the relevant and appropriate information in product title and attributes, such as " *Xiaoxin Air 14*", and "*sky grey*". However, it's may difficult for KOBE model to generate user-cared aspects without the assistance of our proposed posterior distillation module. Our proposed APDT model is able to contain more user-cared information, such as "*easy to carry*", "*very convenient for office*" and "*very smoothly*", since it has learned from the distillation information from reviews (teacher) representation during posterior training phase.

## Model Analysis

**Effect of the Copy Mechanism** To include as many relevant and correctness aspects in the generated product description, the proposed APDT model involves a copy mechanism during the decoder phase. We ablate the copy mechanism from the framework by using only naive transformer decoder to verify its effectiveness. As showed in Table 5, we can witness that the absence of copy mechanism hurts performance of APDT model. The ROUGE-L and BLEU scores decrease from 20.41 to 19.13, and 10.36 to 9.75, respectively. It demonstrates that the copy mechanism plays an important role in achieving strong performance.

| | |
|---|---|
| Product Title | 联想小新Air14高性能游戏办公超轻薄笔记本电脑 Lenovo Xiaoxin Air14 high performance game and office computer with ultra thin and light |
| Attribute sets | **型号**: 小新Air14; **尺寸**: 14英寸; **处理器**: Intel i5-1035G1; **内核数**: 四核; **内存**: DDR4; **显卡**: 独立显卡2GB; **音效系统**: 内置麦克风; **机身材质**: 金属复合材质; **颜色**: 灰色，天空灰(brand name: Xiaoxin Air 14; Size: 14 inch; processor: Intel i5-1035G1; kernel size: four; memory: DDR4; graphics card: SDC 2G; voice: microphone; material: metal composite; color: grey, sky grey ) |
| Reviews (for reference) | R1: 用了一周，觉得很不错。没有卡顿，内存也很大。// R2: 运行速度很快，屏幕清晰，很轻薄，很秀气。// R3: 平时办公用很OK，可以玩游戏，很流畅。携带很方便。//R4: 样子很大气，儿子很喜欢。 R1: used for one week, very nice, no system lag, large capacity. //R2: very fast, clear screen, very thin and light. //R3: it's fine for office job, it also can be used for computer game, very smoothly.//R4: It looks graceful, my son likes it very much. |
| KOBE | 联想小新Air14系列，低调天空灰，外加金属复合材质机身，轻薄系列高性能笔记本电脑。 Xiaoxin Air 14, sky grey with low-profile, plus metal composite material, very thin and high performance nootbook computer. |
| **APDT (ours)** | 联想小新Air14，轻薄笔记本电脑，携带方便，采用四核八线程处理器，游戏体验很流畅，办公更方便。 Xiaoxin Air 14, very thin and light computer, easy to carry, exquisite techniques with four kernel processor, very smoothly to play games, more convenient for office. |

Figure 3: Case study of APDT and baselines on JDPDG dataset.

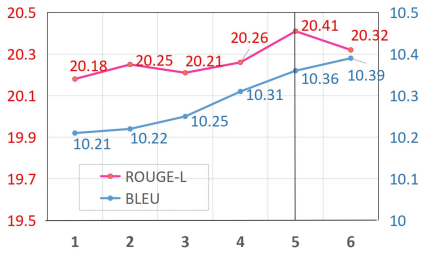
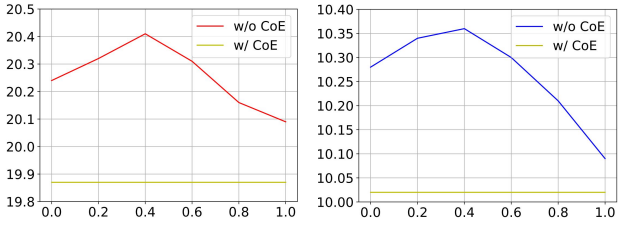



Figure 4: The ROUGE-L (red) and BLEU (blue) scores (%) with different encoder layer of review (teacher) representation in the posterior distillation module.




(a) ROUGE-L score (%). (b) BLEU score (%).

Figure 5: the analysis of $\beta$ in the coherence enhanced function on the validation dataset. Red line in (a) represents the ROUGE-L socre of model with CoE loss, while blue line in (b) stands for the BLEU score of model with CoE loss. Yellow line in both (a) and (b) represent the model without CoE loss.

**Effect of the Posterior Distillation** In our proposed APDT model, posterior distillation can distill user-cared aspect information from review information, allowing the student model to generate the user-cared aspects in the description for long-tailed commodities. From Table 3, the ablating of posterior distillation also leads to a performance drop in the automatic evaluation metrics.

Furthermore, to analyze the distillation effects on product (student) representation, we conduct an experiment to identify which encoder layer that review (teacher) representation distill from. From Figure 4, we observe that the product (student) representation benefits the most from distilling the 5-th or the 6-th encoder layer of the review representation (teacher). In the shallow encoder layer, student representation may not be able to learn well from the user's aspects information. On the other hand, it's easier for student model to mimic from a teacher representation that comes from a deeper layer.

**Effect of the CoE Loss** To evaluate the performance of coherence enhanced negative log-likelihood loss, we explore the impact of using different $\beta$ value in the coherence-enhanced function. As presented in Figure 5, we observe that when $\beta < 0.4$, the increasing of $\beta$ leads to continuous improvement on the BLEU and ROUGE-L metric. However, the performance of our model tends to decrease when $\beta$ continues rising. Therefore, overemphasis on the coherence-enhanced function will finally risk introducing inconsistent aspects information into the final description sentences.

## Conclusion

In this paper, we propose an adaptive posterior distillation method for product description generation task. This method enables our Transformer-based model to utilize customer reviews and incorporate user-cared aspects into product description, especially for the long-tailed commodities. To better evaluate our proposed approach, we also construct a new Chinese product description dataset CPDG, and then present an adaptive posterior distillation method, which can distill user-cared aspects to the product description generation process. Extensive experiments conducted on our proposed dataset show that our proposed method could achieve better performance than baseline models. In future work, we plan to further investigate the proposed model with question-answering information, and then extend our approach to a multi-task framework, which is capable to handle a joint user intent recognition task.

## Acknowledgements

The authors would like to thank Hengyi Cai from Institute of Computing Technology, Chinese Academy of Sciences, and the anonymous reviewers for their constructive comments and suggestions. This work was partially supported by the National Key R&D Program of China under Grants No. 2019AAA0105200, 2016QY02D0405, the Beijing Academy of Artificial Intelligence (BAAI) (No. BAAI2020ZJ0303), the National Natural Science Foundation of China (NSFC) (No. 61722211, 61773362, 61872338, 61906180).

## References

Cai, H.; Chen, H.; Song, Y.; Zhao, X.; and Yin, D. 2020. Exemplar Guided Neural Dialogue Generation. In *IJCAI*, 3601–3607.

Carenini, G.; and Moore, J. D. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence* 170(11): 925–952.

Chen, M.; Liu, R.; Shen, L.; Yuan, S.; Zhou, J.; Wu, Y.; He, X.; and Zhou, B. 2020. The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service. In *LREC*, 459–466.

Chen, Q.; Lin, J.; Zhang, Y.; Yang, H.; Zhou, J.; and Tang, J. 2019a. Towards Knowledge-Based Personalized Product Description Generation in E-commerce. In *SIGKDD*, 3040–3050.

Chen, W.; Huang, P.; Xu, J.; Guo, X.; Guo, C.; Sun, F.; Li, C.; Pfadler, A.; Zhao, H.; and Zhao, B. 2019b. POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion. In *SIGKDD*, 2662–2670.

Ding, T.; and Pan, S. 2016. Personalized Emphasis Framing for Persuasive Message Generation. In *EMNLP*, 1432–1441.

Elad, G.; Guy, I.; Novgorodov, S.; Kimelfeld, B.; and Radinsky, K. 2019. Learning to Generate Personalized Product Descriptions. In *CIKM*, 389–398.

Fabbri, A.; Li, I.; She, T.; Li, S.; and Radev, D. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *ACL*, 1074–1084.

Feng, S.; Chen, H.; Li, K.; and Yin, D. 2020a. PosteriorGAN: Towards Informative and Coherent Response Generation with Posterior Generative Adversarial Network. In *AAAI*, 7708–7715.

Feng, S.; Ren, X.; Chen, H.; Sun, B.; Li, K.; and Sun, X. 2020b. Regularizing Dialogue Generation by Imitating Implicit Scenarios. In *EMNLP*, 6592–6604.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5): 378.

Gong, Y.; Luo, X.; Zhu, K. Q.; Ou, W.; Li, Z.; and Duan, L. 2019. Automatic generation of chinese short product titles for mobile display. In *AAAI*, volume 33, 9460–9465.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* .

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* .

Krishna, K.; Murhekar, A.; Sharma, S.; and Srinivasan, B. V. 2018. Vocabulary Tailored Summary Generation. In *COLING*, 795–805.

Li, H.; Yuan, P.; Xu, S.; Wu, Y.; He, X.; and Zhou, B. 2020. Aspect-Aware Multimodal Summarization for Chinese E-Commerce Products. In *AAAI*, 8188–8195.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, Y.; and Lapata, M. 2019. Hierarchical Transformers for Multi-Document Summarization. In *ACL*, 5070–5081. Florence, Italy: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 8024–8035.

Pecar, S. 2018. Towards opinion summarization of customer reviews. In *Proceedings of ACL 2018, Student Research Workshop*, 1–8.

Reichelt, M.; Kämmerer, F.; Niegemann, H. M.; and Zander, S. 2014. Talk to me personally: Personalization of language style in computer-based learning. *Computers in Human behavior* 35: 199–210.

Roy, R. S.; Padmakumar, A.; Jeganathan, G. P.; and Kumaraguru, P. 2015. Automated linguistic personalization of targeted marketing messages mining user-generated text on social media. In *CICLing*, 203–224. Springer.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*, 1073–1083.

Shahidi, H.; Li, M.; and Lin, J. 2020. Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data. In *ACL*, 3864–3870.

Shen, L.; and Feng, Y. 2020. CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation. In *ACL*, 556–566.

Shen, L.; Feng, Y.; and Zhan, H. 2019. Modeling Semantic Relationship in Multi-turn Conversations with Hierarchical Latent Variables. In *ACL*, 5497–5502.

Shen, L.; Guo, X.; and Chen, M. 2020. Compose Like Humans: Jointly Improving the Coherence and Novelty for Modern Chinese Poetry Generation. *arXiv preprint arXiv:2005.01556* .

Shen, L.; Zhan, H.; Shen, X.; and Feng, Y. 2021. Learning to Select Context in a Hierarchical and Global Perspective for Open-domain Dialogue Generation. *arXiv preprint arXiv:2102.09282* .

Sun, F.; Jiang, P.; Sun, H.; Pei, C.; Ou, W.; and Wang, X. 2018. Multi-source pointer network for product title summarization. In *CIKM*, 7–16. ACM.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.

Wang, J.; Hou, Y.; Liu, J.; Cao, Y.; and Lin, C.-Y. 2017. A statistical framework for product description generation. In *IJCNLP*, 187–192.

Wang, J.; Tian, J.; Qiu, L.; Li, S.; Lang, J.; Si, L.; and Lan, M. 2018. A multi-task learning approach for improving product title compression with user search log data. In *AAAI*.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* .

Zander, S.; Reichelt, M.; Wetzel, S.; et al. 2015. Does Personalisation Promote Learners' Attention? An Eye-Tracking Study. *Frontline Learning Research* 3(4): 1–13.

Zhan, H.; Zhang, H.; Chen, H.; Shen, L.; Lan, Y.; Ding, Z.; and Yin, D. 2020. User-Inspired Posterior Network for Recommendation Reason Generation. In *SIGIR*, 1937–1940.

Zhang, J.; Zou, P.; Li, Z.; Wan, Y.; Pan, X.; Gong, Y.; and Philip, S. Y. 2019. Multi-Modal Generative Adversarial Network for Short Product Title Generation in Mobile E-Commerce. In *NAACL-HLT*, 64–72.