TemPEST: Soft Template-Based Personalized EDM Subject Generation through Collaborative Summarization

Yu-Hsiu Chen, Pin-Yu Chen, Hong-Han Shuai, Wen-Chih Peng
National Chiao Tung University, Hsinchu, Taiwan
{yhchen.cm06g, pinyu.eed04, hhshuai}@nctu.edu.tw, wcpeng@g2.nctu.edu.tw

Abstract

We address personalized Electronic Direct Mail (EDM) subject generation, which generates an attractive subject line for a product description according to user's preference on different contents or writing styles. Generating personalized EDM subjects has a few notable differences from generating text summaries. The subject has to be not only faithful to the description itself but also attractive to increase the click-through rate. Moreover, different users may have different preferences over the styles of topics. We propose a novel personalized EDM subject generation model named Soft Template-based Personalized EDM Subject Generator (TemPEST) to consider the aforementioned users' characteristics when generating subjects, which contains a soft template-based selective encoder network, a user rating encoder network, a summary decoder network and a rating decoder. Experimental results indicate that TemPEST is able to generate personalized topics and also effectively perform recommending rating reconstruction.

Introduction

With the rapid growth of e-commerce, a variety of advertising strategies have been implemented to convert prospects into paying customers. One of the effective advertising ways is Electronic Direct Mail (EDM). According to the e-commerce conversion rate benchmark with more than 10,000 online stores conducted by Compass, EDM is the highest converting channel for almost every product type.[1] Nevertheless, the conversion rate of EDM is still low, i.e., 5.32% median conversion rate in 2018, which still leaves a lot of rooms for improvement. To boost the conversion rate, personalized subject lines have been proved to be effective for increasing open/conversion rates. The report from Experian Marketing Services indicates that emails with personalized subject lines create a boost of 37% open rates for emails compared to emails sent during the same period with non-personalized subject lines.[2]

Table 1 illustrates the motivation with three EDM subject lines. Given the product description of a Hong-Kong traveling package, all subject lines shown in the table can be used as the subject of EDM since they can all represent the meaning in the article. User 1 clicks into the one with an exaggerating subject line, whereas user 3 likes a more general one. Therefore, in this paper, we argue that different styles of EDM subjects can be personalized to attract different users.

Table 1: Example of the article subject generation. Personalized subject summarization is motivated by the fact that different users are inclined to click in different subjects from the same articles, according to their preference on different styles.

User 1	【急速挑戰】香港驚險瀑布湍流攀爬挑戰 (The Top Speed Challenge: Thrilling Waterfall and Rapid Stream Climbing Challenge in Hong Kong)
User 2	【香港必去】屏南石澗瀑布自然風光 (Hong Kong Must-See: Ping Nam Waterfall Landscapes)
User 3	【香港人氣景點】屏南石澗一日遊 (Hong Kong Popular Attractions: One Day Trip to Ping Nam Waterfall)

However, current personalized EDM subjects require either human editors or are based on predefined rules. With the advance of deep learning, it is promising to use a data-driven approach for personalizing the EDM subjects. One of the possible solutions is to use article summarization (Cao et al. 2018b; Gao et al. 2019), which aims to summarize the content with a few sentences. Alternatively, personalized review generation (Ni and McAuley 2018; Li, Li, and Zong 2019) is another method that seeks to help users in their choices or the understanding of the recommendation. Nonetheless, article summarization does not take the attractiveness of EDM subject lines into consideration. Furthermore, personalized review generation cannot be applied to generate EDM subject lines since the review generation is conditioned on a predefined aspect representation words list which is unavailable for EDM subject generation.

Therefore, this paper addresses the personalized subject generation task which has not been discussed in previous research. Given the article content, personalized subject generation task aims to summarize the main idea and to make the style match with the users' preference simultane-

[1] http://bit.ly/2lsVEpI

[2] http://bit.ly/2kc7OD7

ously. Consequently, it is related to both summarization and recommendation problems. As such, we exploit interactive users' clicks as ratings and include collaborative filtering to take both rating and content information into consideration. Specifically, we propose the *Soft Template-based Personalized EDM Subject Generation model* (TemPEST), which consists of two bi-directional selective encoders and an ordinary RNN decoder. The first encoder, namely *Template-aware Sequence Encoder* (TSE), jointly selects important information from the source article and its corresponding template to assist with better representations. Moreover, the second, namely *User-aware Sequence Encoder* (USE), utilizes the styles of users' preferable subjects and the article representation from the first encoder to generate user-specific article representation. Afterward, an RNN decoder is used to construct the target word distribution from the latent representation generated from USE. In other words, we combine the user's click-through subjects to generate the representation consisting of his/her preferable style by TSE and USE and generate the EDM subject by the RNN decoder.

Besides using the preferable subjects from users' browsing history, we propose to use a collaborative recommender to guide the model for generating attractive subject lines. Therefore, a gating layer is used for integrating user rating representation and user-specific article representation from the second User-aware Sequence Encoder above. A two-layer decoder simultaneously reads the gated item embedding and its similar item embedding, which is captured by a neighbor-attention module, and finally reconstructs ratings. The rating decoder then reconstructs ratings and returns a reconstruction loss to guide a better user rating representation. It is worth noting that the proposed model is capable of recommending top-k items with a better rating reconstruction.

To validate our model, we collect a new dataset comprising 17617 products including titles with corresponding descriptions and 278876 ratings from 21379 users. Experimental results on quantitative and qualitative evaluation manifest that TemPEST achieves state-of-the-art performance on personalized subject generation. A user study with 81 users shows that the subject lines generated by TemPEST are 59.07% more attractive than that generated by the state-of-the-art method. Moreover, the similarity score between subjects and users' preferable articles outperforms the real subject by 6.8%, while the relevance of TemPEST is comparable to the state-of-the-art template-based subject generator. The contributions are summarized as follows:

- We propose a personalized Seq2Seq-based model named Soft Template-based Personalized EDM Subject Generation Network (TemPEST) to generate the personalized EDM subjects. To the best of our knowledge, this is the first work exploiting a data-driven approach for personalized EDM subject generation.
- The proposed model adopts selective gates and collaborative filtering mechanism to consider users' preference of subject styles, while the selective encoding is also able to perform collaborative filtering for the recommendation system.
- Experimental results show that, in terms of subject generation, our model demonstrates promising results in relevance and preference from both qualitative and quantitative measurements. In addition, the proposed TemPEST is also able to recommend preferable items to users.

Preliminaries

Related Work

Text summarization can be categorized into extractive and abstractive methods. Extractive methods (Narayan, Cohen, and Lapata 2018; Jadhav and Rajan 2018) select sentences from articles and combine them into a paragraph, while abstractive methods (Cao et al. 2018b; Gao et al. 2019) rewrite the summaries, which may not be featured in the source text. Currently, sequence-to-sequence (Seq2Seq) frameworks (Chopra, Auli, and Rush 2016; Nallapati et al. 2016) have become the mainstream among the abstractive models due to its effectiveness. On the other hand, template-based summarization is another effective approach to abstractive summarization. Given a set of templates, hard template-based models are trained to extract and populate key snippets into the templates to form the final summaries (Zhou and Hovy 2004), which guarantee concise and coherent summaries. However, template-based summarization is usually time-consuming and lacking in domain knowledge to create all template manually. In contrast, soft template-based models (Cao et al. 2018a; Wang, Quan, and Wang 2019) select the summaries of a specific training article as a soft template. Nevertheless, the role of users' preference style remains unexplored in subject generation. The idea of personalized review generation mostly aims to generate different aspects of reviews (Li, Li, and Zong 2019; Ni and McAuley 2018; Liu et al. 2018), which either exploits users' past reviews as the additional input or utilizes users' annotation/highlight words to capture different aspects as conditioning attributes. Yet, this kind of studies need a predefined representative word list for inferred aspects on a certain dataset and is unrealistic to list all words manually.

To correctly model the users' preference on the EDM subject, a personalized recommendation system is required since it captures the user's preference from his/her historical interaction with items and recommends some promising items. With the advance of deep learning, autoencoders have become popular to model the users' preference (Wang, Wang, and Yeung 2015; Wu et al. 2016; Li and She 2017; Ma et al. 2019). Wu et al. (2016) apply the idea of denoising autoencoder into recommender system and reconstruct the user's implicit feedback from the user's latent representation. Li and She (2017) use variational autoencoder to learn deep latent representation from item content, and jointly model the generation of latent content and user ratings. A recent line of studies incorporates user-item ratings for generation task. For instance, Vo and Soh (2018) create new (non-existent) item feature vectors to satisfy groups of users with different preferences. Moreover, Truong and Lauw (2019) input both reviews and ratings to generate a more holistic representation of a review guided by the recommender, which generates a document according to rat-

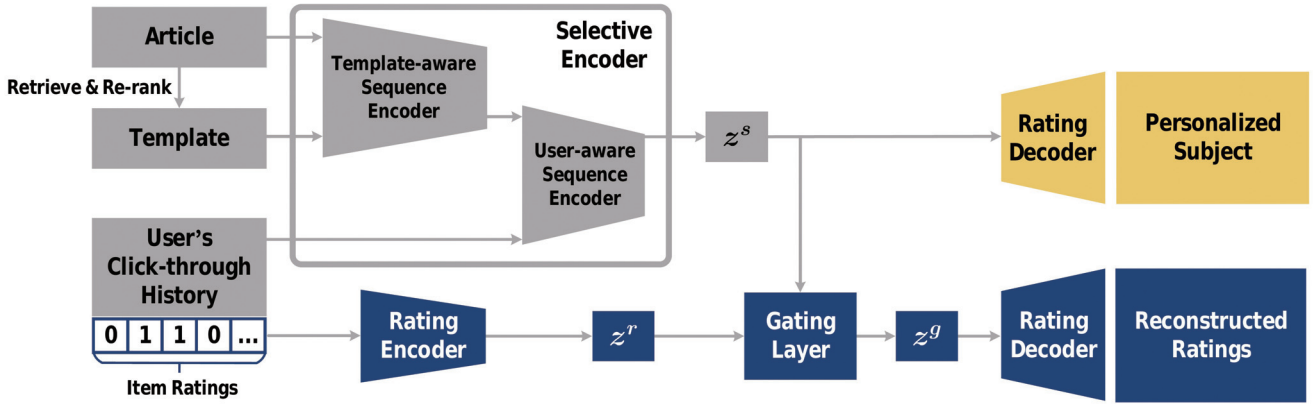

Figure 1: The overview architecture of our proposed model. The grey part indicates the selective encoder of TemPEST including Template-aware Sequence Encoder (TSE) and User-aware Sequence Encoder (USE). The yellow part is the RNN summary decoder and the blue part is the recommender gated attentive autoencoder. The *Item Ratings* part on the lower left indicates the binary ratings constructed by the above click-through history.

ings. However, to the best of our knowledge, this is the first work that utilizes product descriptions and user ratings to generate a personalized subject and boost the recommender system at the same time.

Problem Formulation

Suppose we have a corpus D with m article-template-subject triples a-t-s, and each triple contains an article a, a template t and a subject s. Article a consists of l words as $\{a_1, a_2, ..., a_l\}$, where $a_i \in D$. Template t and subject s consist of $p \leq l, q \leq l$ words as $\{t_1, t_2, ..., t_p\}$ and $\{s_1, s_2, ..., s_q\}$ individually, where $t_i, s_i \in D$. In addition, the user preferences are presented by an n-by-m binary matrix $\mathbf{R}$. $\mathbf{R}_{ij} = 1$ if user i has clicked into the item j.

The personalized subject generation task intends to: 1) retrieve template t given article a, and then further summarize a subject $\hat{s}$ given a and t by attending to user u's preference item subjects, 2) predict the rest of ratings in $\hat{\mathbf{R}}$ given a part of the ratings in $\mathbf{R}$.

Proposed Model

To generate personalized EDM subjects, there are three challenges required to be addressed. First, different users prefer different styles of headlines, while it is difficult to explicitly define all styles. One possible solution is to select a subject line among users' clicking history and exploit template-based summarization method to generate a personalized subject. However, this approach is limited since only one clicked subject line is used. Second, to summarize the article, articles are encoded into a latent vector. Nevertheless, the style and content are not disentangled in the latent space. Therefore, it is necessary to design a loss for disentangling the style and content so that the summarization only changes the style and preserves the content. Third, the clicking history of users may be sparse for new users, which leads to the "cold start" problem. Therefore, to overcome these issues, it requires combining the generation and recommendation elegantly.

Keep the goals in mind, in this paper, we propose a new framework including three key modules: **Retrieve and re-rank**, **TemPEST**, and **Gated attentive autoencoder**. **Retrieve and re-rank** aims to return a few candidate templates from the training corpus and re-rank by semantic relationship to identify the best one. To jointly address the first and second issues, **TemPEST** mutually selects important information from the source article and template to generate an article representation of summarization. Afterward, it further inputs the users' preference subject contents and outputs an enhanced representation attending users' preference style. Finally, with generated representation and user-item ratings, the **Gated attentive autoencoder** learns fused hidden representation to address the third issue. The overview model architecture is illustrated as shown in Figure 1.

Retrieve and Re-rank

The module intends to find several candidate templates from the training corpus and select the closest one based on both word frequency and semantic distance. We assume that similar sentences contain similar summary patterns. Given a product description, the widely-used information retrieval library Lucene[3] is used to retrieve a set of similar paragraphs and use their summaries as the candidate templates. We use the default settings of Lucene as (Cao et al. 2018a) to build the information retrieval system for index and search. For each paragraph, we select top 30 searching results as candidate templates of a certain article.

To measure the deep semantic relationship between initial search candidates and the ground truth subject to select the best template, we use the embedding space cosine similarity between two titles to identify the best one since the previous retrieve method can only calculate the word matching

[3]https://lucene.apache.org

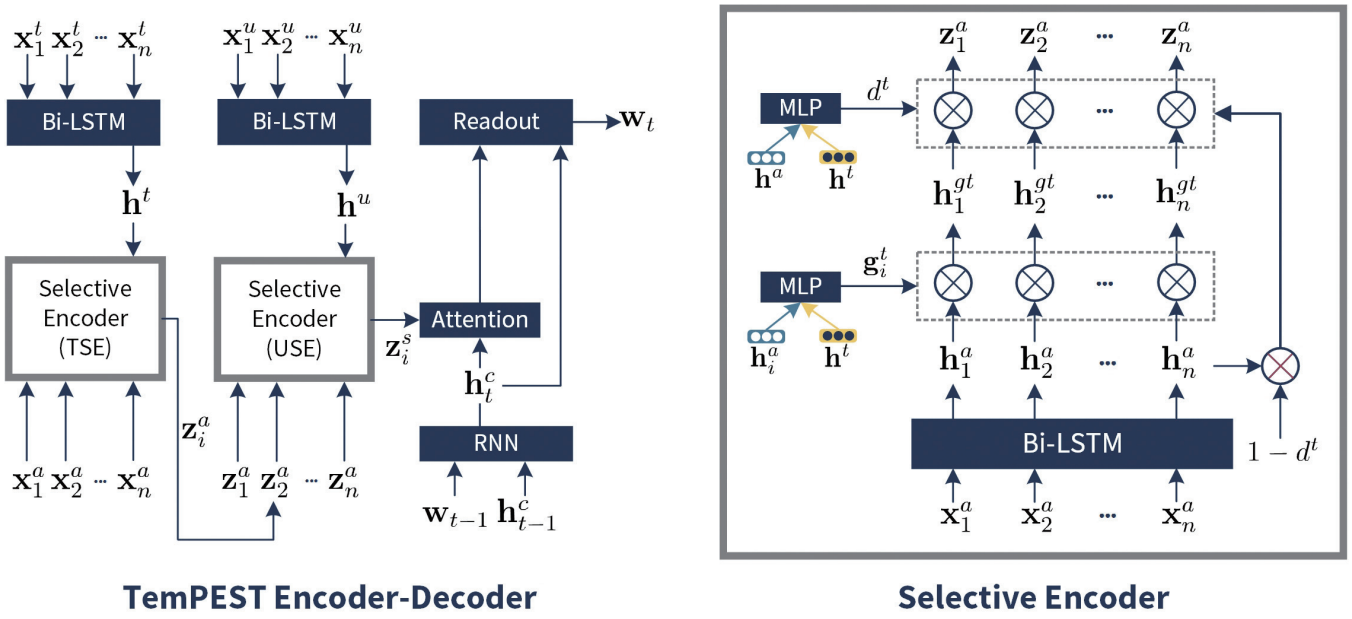

Figure 2: The TemPEST selective encoder-decoder architecture. The architecture on the left is the overall structure of TemPEST. The structure of selective encoders bounded by grey boxes are illustrated on the right. Note that the notation on the right depicts the first Template-aware Sequence Encoder (TSE).

similarities. We train a Word2Vec embedding (Mikolov et al. 2013a) with Wikipedia. A similarity score s is calculated for a template representation $\mathbf{z}_i^t$ and a gold subject representation z^s:

$$s_i = (\mathbf{z}_i^t)^{\mathrm{T}} \mathbf{z}^s.$$

The subject with the highest similarity score is retrieved as the template, i.e., $\arg\max_i \ s$.

It is worth noting that the previous re-rank settings usually use ROUGE score (Lin 2004) to evaluate the saliency of a candidate template. However, using a ROUGE score captures only word-level similarity. For instance, when generating EDM subject lines for the product of tour packages, subjects contain "one day tour" returns high ROUGE score since similar words often appear in subjects. However, our goal is to focus on capturing important words to contribute more to the semantic meanings. Therefore, we use Word2Vec similarity (Vulić and Moens 2015) to select our desired template.

TemPEST

Inspired by the research in selective mechanism (Zhou et al. 2017; Wang, Quan, and Wang 2019), we propose a novel soft **Tem**plate-based **P**ersonalized **E**DM **S**ubject generation (TemPEST) module for personalized summarization. The core idea behind TemPEST is to integrate the templates and users' preference style sentences to generate a personalized stylish article representation and subject generation. Figure 2 illustrates the architecture of TemPEST, which comprises two encoders and one decoder.

For the first *Template-aware Sequence Encoder* (TSE), we adopt two gates to mutually select important information from a given source article and its template. The goal of first Template-to-Article gate is to use the template to filter the article representation. For each time step i, the selective gate takes the representation of template $\mathbf{h}^t$ and article's BiLSTM hidden state $\mathbf{h}_i^a$ as inputs to output a template gate vector $\mathbf{g}_i^t$ to select $\mathbf{h}_i^a$:

$$\mathbf{g}_i^t = \sigma(\mathbf{W}_{ah}\mathbf{h}_i^a + \mathbf{W}_{th}\mathbf{h}^t + \mathbf{b}_a),$$

$$\mathbf{h}_i^{gt} = \mathbf{h}_i^a \odot \mathbf{g}_i^t,$$

where $\mathbf{h}^a$ is the concatenation of the last forward hidden state $\overrightarrow{\mathbf{h}_n^t}$ and the first backward hidden state $\overleftarrow{\mathbf{h}_1^t}$ of the template. $\mathbf{W}_{ah}$, $\mathbf{W}_{th}$, and $\mathbf{b}_a$ are learnable parameters, σ denotes sigmoid activation function, and $\odot$ is element-wise multiplication.

On top of that, the role of second Article-to-Template gate is to decide the proportion of $\mathbf{h}^{gt}$ in the final representation. Therefore, a confidence degree d^t is learned to decide the weight of $\mathbf{h}_i^{gt}$ which is computed by:

$$d^t = \sigma((\mathbf{h}^a)^{\mathrm{T}}\mathbf{W}_d\mathbf{h}^t + \mathbf{b}_{dt}),$$

where article representation $\mathbf{h}^a$ is obtained in a similar way as $\mathbf{h}^t$. The article representation is the weighted sum of $\mathbf{h}_i^a$ and $\mathbf{h}_i^{gt}$:

$$\mathbf{z}_i^a = d^t\mathbf{h}_i^{gt} + (1 - d^t)\mathbf{h}_i^a.$$

To further manipulate the style of the article summaries above, we add another two gates in the second *User-aware Sequence Encoder* (USE) similar to the above setting. The first User-to-Summary gate filters the above summarized article representation and the second Summary-to-User gate

decides the proportion. The final user-specific article representation is denoted $\mathbf{z}_i^s$ as follows:

$$\mathbf{z}_i^s = d^u \mathbf{h}_i^g + (1 - d^u)\mathbf{h}_i^s.$$

After selecting important information by using the previous encoder, RNN decoder is used to generate the subject line. At each time step t, the decoder reads the previous word embedding $\mathbf{w}_{t-1}$ and hidden state $\mathbf{h}_{t-1}^c$ generated in the previous step to compute the new hidden state:

$$\mathbf{h}_t^c = \mathrm{RNN}(\mathbf{w}_{t-1}, \mathbf{h}_{t-1}^c).$$

Let $\mathbf{h}^a$ denote the original article representation and is initialized as the hidden state. The context vector $\mathbf{c}_t$ for current time step t is computed via the concatenate attention mechanism (Luong, Pham, and Manning 2015) between $\mathbf{h}_t^c$ and final personalized article representation $\mathbf{z}^s$:

$$\varepsilon_{t,i} = (\mathbf{z}_i^s)^{\mathrm{T}} \mathbf{W}_c \mathbf{h}_t^c,$$

$$\alpha_{t,i} = \frac{\exp\left(\varepsilon_{t,i}\right)}{\sum_{i=1}^{M} \exp\left(\varepsilon_{t,i}\right)},$$

$$\mathbf{c}_t = \sum_{i=1}^{L} \alpha_{t,i} \mathbf{z}_i^s.$$

Afterward, a concatenation layer is exploited to combine the hidden state $\mathbf{h}_t^c$ and the context vector $\mathbf{c}_t$ into a new overall hidden state $\mathbf{h}_t^o$:

$$\mathbf{h}_t^o = \tanh(\mathbf{W}_{ho}[\mathbf{c}_t; \mathbf{h}_t^c]),$$

$\mathbf{h}_t^o$ is fed through a softmax layer to output the target word distribution:

$$p(\mathbf{w}_t \mid \mathbf{w}_1, ..., \mathbf{w}_{t-1}) = \mathrm{softmax}(\mathbf{W}_p \mathbf{h}_t^o).$$

In summary, the loss function of TemPEST consists of three parts. To learn the generation of subjects, we minimize the negative log-likelihood between generated subject $\mathbf{w}$ and gold summary $\mathbf{w}^*$:

$$\mathcal{L}_s = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{L} \log p(\mathbf{w}_j^{*(i)} | \mathbf{w}_{j-1}^{(i)}, \mathbf{x}^{a(i)}, \mathbf{x}^{t(i)}),$$

where L is the length of the gold summary, θ is the parameter trained in our model, and $\mathbf{x}^a$ and $\mathbf{x}^t$ indicate the source article and the template, respectively.

To learn the style of the template, we also minimize the negative log-likelihood between generated subject $\mathbf{w}$ and the template $\mathbf{w}^t$:

$$\mathcal{L}_t = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{L} \log p(\mathbf{w}_j^{t(i)} | \mathbf{w}_{j-1}^{(i)}, \mathbf{x}^{a(i)}, \mathbf{x}^{t(i)}).$$

$\mathbf{w}_j^{t(i)}$ denotes the i-th word and the j-th sentence in template $\mathbf{w}^t$.

For user-aware subject generation, the loss function between generated subject $\mathbf{w}$ and user template $\mathbf{w}^u$ is:

$$\mathcal{L}_u = -\frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{L} \log p(\mathbf{w}_j^{u(i)} | \mathbf{w}_{j-1}^{(i)}, \mathbf{x}^{a(i)}, \mathbf{x}^{t(i)}, \mathbf{x}^{u(i)}).$$

where $\mathbf{w}_j^{u(i)}$ is the i-th word and j-th sentence in user template $\mathbf{w}^u$ collected from users clicked through subjects.

Gated Attentive Autoencoder

To address the third challenge, we propose a gated attentive autoencoder based on Ma et al. (2019). A stacked autoencoder is applied to encode users' binary ratings on certain item k into item's rating representation $\mathbf{z}_k^r$. The encoder part is:

$$\mathbf{z}_k^r = \tanh\left(\mathbf{W}_{e2} \tanh\left(\mathbf{W}_{e1} \mathbf{r}_k + \mathbf{b}_{e1}\right) + \mathbf{b}_{e2}\right),$$

where $\mathbf{W}_{e1}$, $\mathbf{W}_{e2}$, $\mathbf{b}_{e1}$, and $\mathbf{b}_{e2}$ are parameters in the 2-layer rating encoder. Item k's latent representation $\mathbf{z}^{s(k)}$ is obtained from TemPEST encoder, and is the average of $\{z_j^s\}_{j=1}^L$. A neural gating layer is used to combine item hidden representations from two data sources (i.e., $\mathbf{z}_k^r$ and $\mathbf{z}^{s(k)}$), and we get fused item hidden representation $\mathbf{z}_k^g$:

$$\mathbf{G} = \sigma(\mathbf{W}_{g1} \mathbf{z}_k^r + \mathbf{W}_{g2} \mathbf{z}^{s(k)} + \mathbf{b}_g),$$

$$\mathbf{z}_k^g = \mathbf{G} \odot \mathbf{z}_k^r + (1 - \mathbf{G}) \odot \mathbf{z}^{s(k)}.$$

Items similar with certain item k are also captured by the neighbor-attention module. The similarity calculation is based on items' common users in preprocessing stage. The neighbor-embedding $\mathbf{z}_k^n$ of item k with the set of neighbors, N_k, of item k is defined by:

$$s_{kj} = \tanh\left(\mathbf{z}_k^{g\top} \mathbf{W}_n \mathbf{z}_j^g\right), \forall j \in N_k,$$

$$\mathbf{a}_k = \mathrm{softmax}(\mathbf{s}_k),$$

$$\mathbf{z}_k^n = \sum_{j \in N_k} a_{kj} \mathbf{z}_j^g.$$

The rating decoder simultaneously reads fused item hidden representation $\mathbf{z}_k^g$ and its similar neighbor-embedding $\mathbf{z}_k^n$ as follows.

$$\mathbf{z}_k^{g^*} = \tanh\left(\mathbf{W}_{d1} \mathbf{z}_k^g + \mathbf{b}_{d1}\right),$$

$$\mathbf{z}_k^{n^*} = \tanh\left(\mathbf{W}_{d1} \mathbf{z}_k^n + \mathbf{b}_{d1}\right),$$

$$\hat{\mathbf{r}}_k = \sigma(\mathbf{W}_{d2} \mathbf{z}_k^{g^*} + \mathbf{W}_{d2} \mathbf{z}_k^{n^*} + \mathbf{b}_{d2}).$$

We use a squared error for the loss of reconstruction ratings:

$$\mathcal{L}_r = \sum_{i=1}^{n} \sum_{j=1}^{m} ||\mathbf{C}_{ij}(\mathbf{R}_{ij} - \hat{\mathbf{R}}_{ij})||_2^2 + \lambda ||\mathbf{W}_*||_F^2,$$

where λ is the regularization parameter, and $||\cdot||_F$ is Frobenius norm of matrices. The confidence $\mathbf{C}_{ij}$ is defined as:

$$\mathbf{C}_{ij} = \begin{cases} \rho, & \text{if } \mathbf{R}_{ij} = 1 \\ 1, & \text{otherwise.} \end{cases}$$

The total loss of the proposed model is shown as follows:

$$\mathcal{L} = \mathcal{L}_s + \lambda_t \mathcal{L}_t + \lambda_u \mathcal{L}_u + \lambda_r \mathcal{L}_r.$$

Table 2: Dataset statistics for *KKday*.

# products	17617	# title	17617
# users	21379	# word/title	12.8
# item	5365	#word/article	155.2
# item/user	13.04	# vocabulary	71625

Experimental Results

Dataset

Since there is no public dataset for personalized subject summarization, we collect a new one named *KKday* from the well-known travel experience e-commerce platform in Asia that sales tour packages.[4] The raw data contains the paired tuple (article, subject) of tour package DMs, together with the ID list of users who click the subject. Statistics of *KKday* are summarized in Table 2. Specifically, there are 17617 products associated with the EDM subject and a short paragraph of the product description. Each description contains 885 Traditional Chinese characters on average and is tokenized by CKIP (Sproat and Emerson 2003).

For the user clicking histories, we collect the data from December 2018 to July 2019, which contains 26,662,557 user-item interaction with 1,271,297 users and 9,702 items. We observe that users interact with items sequentially between a small period. This is usually the situation that they are first attracted by the subject of a tour package and continue to browse related packages, which may have the same tour destination with the first one. To approximately collect items whose subjects are most attractive to the user, we simply take the first item viewed by the user for each viewing sequence as the ground truth. After filtering out users with less than 10 items viewed and items without descriptions, there are 278,876 records with 21,379 users and 5,365 items.

For summarization task, we randomly split the dataset into 14095 products for training, 1761 for testing and 1761 for validation. The input vocabularies are collected from the training data, which have 71625 words in total. For recommendation task, we randomly select 80% of each user's records for training, and the remaining data are for testing.

Evaluation Metrics

For relevance between generated and reference sentences, we adopt BLEU (Kishore Papineni and Zhu 2002) and ROUGE (Lin 2004). BLEU is a well-known metric for evaluating the quality of the generated text, which has been widely used for machine translation and image captioning. We use smoothed BLEU (Lin and Och 2004) and report the results of BLEU from 1 to 4. In contrast, ROUGE is a commonly-used metric for text summarization. We report the F-measure, which is the geometric mean of the precision and recall of ROUGE-1 and ROUGE-L.

Besides, for styling relevance between generated sentences and user-specific sentences, we use the Word2Vec similarity (Mikolov et al. 2013a; 2013b) to measure the similarity between generated subject and source article. With the linear linguistic regularities property in Word2Vec embedding space (Vulić and Moens 2015), the semantic relation is maintained by simple addition. Following the settings in Vulić and Moens (2015), all the words in a subject and an article are converted to word embedding by Word2Vec model, and we take average among their word embedding to make a subject embedding or an article embedding. The Word2Vec is implemented with *gensim* and pretrained on the latest Chinese Wikipedia. Finally, we evaluate the performance of the recommender system by *Recall@K* and *NDCG@K*. Recall@K indicates the percentage of the recommended items among items relevant to the user. NDCG@K is short for Normalized Discounted Cumulative Gain at K, which takes the position of the recommended items' order into account.

[4] https://www.kkday.com

Table 3: BLEU and ROUGE F1 scores of all methods on the test set. B in the table denotes BLEU and R in the table denotes ROUGE. The best method is marked as **bold**.

Models	B-1	B-2	B-3	B-4	R-1	R-L
Lead-1	6.16	2.20	1.29	0.78	46.93	6.34
S2S+attn	20.59	4.49	2.59	1.55	50.25	5.65
BiSET	22.06	4.54	2.61	1.52	55.54	**8.70**
TemPEST	**24.36**	**5.31**	**3.15**	**1.97**	**65.47**	8.33

Baselines

Here, we compare our model with several methods which are popular in text summarization as follows:

- *Lead*-1 is an extractive approach which selects the first sentence in review as a summary.
- *S2S+Att* is a sequence-to-sequence model with attention implemented by the OpenNMT (Klein et al. 2017).
- *BiSET* (Wang, Quan, and Wang 2019) adopts a selective network to bidirectionally select important information from template and article, which obtains state-of-the-art in template-based summarization.
- *TemPEST* is the proposed model. We use two-layer Bi-LSTM for both TSE and USE network, and the hidden state size of 500. The learning rate and dropout rate are set to 0.001 and 0.3 respectively.

Results and Discussions

Qualitative Results

BLEU and ROUGE Scores Table 3 presents the results of BLEU scores and ROUGE scores, which manifests that TemPEST consistently outperforms the baselines in terms of BLEU scores, i.e., TemPEST synthesizes subjects that are closer to the ground truth references. For ROUGE score, all ROUGE-1 scores are pretty high since subjects in *KKday* contain a lot of proper nouns like place names to describe the products. Therefore, the extractive method *Lead*-1 performs relatively well comparing to other summarization datasets. Note that the ROUGE-L scores are comparatively low in this task since the subjects in *KKday* are succinct and the place names do not need to appear consecutively in certain order to be informative. Hence, *S2S+Att* performs even poorer than *Lead*-1 in terms of ROUGE-L, which also justify the reason why we adopt template-based summarization methods. For *BiSET*, it significantly improves *S2S+Att* and *Lead*-1, which shows the selective mechanism to filter the input plays an important role for summarization.

Finally, our proposed TemPEST is slightly less than *BiSET* in ROUGE-L. Since the evaluation metric has only a gold summary with a certain style. Imagine that we simply rewrite subjects from the ground truth answer and make

Table 4: Word2Vec cosine similarity metrics of all methods. The best performing method is marked as **bold** while the second best performing method is marked as underlined.

Models	Summary similarity	User similarity
S2S+Att	0.551	0.409
BiSET	0.580	0.393
TemPEST	0.592	**0.456**
Reference	**0.635**	0.427

Table 5: The recommendation performance comparison of all methods in terms of $Recall@10$ and $NDCG@10$. The best performance is marked as **bold**.

Models	Recall@10	NDCG@10
CDAE	0.0629	0.0199
GATE	0.1871	0.0619
TemPEST	**0.1921**	**0.0632**

them transfer to another style preserving the informativeness. The ROUGE metric cannot capture this kind of information well and hence the scores will decrease. In short, TemPEST attempts to strike a balance between the personalized styling and the faithfulness. Since there is no way that we can generate a sentence looking just the same as ground truth answer but in different styles, we adopt another kind of evaluation metric to estimate the effect of personalizing.

Word2Vec Similarity In addition to the comparison between the ground truth subject and generated subject at the word level, we want to further estimate them at the semantic level. Table 4 shows the summary similarity, i.e., the cosine similarity between source article and different subject lines, for different approaches. The similarity between the reference subject and source article is the highest which is to be expected. The results show that, even though TemPEST is not the best in word level, the semantic similarity is still close to the given source article and outperforms BiSET. Therefore, synonyms can be used to transfer the styles without changing its meaning.

To evaluate whether the generated subject is close to the user's preference, we calculate the average embedded similarity scores between the user's clicked through subjects and generated subjects. The proposed TemPEST outperforms BiSET by 16% and is even better than the ground truth subjects. *S2S+Att* slightly surpasses *BiSET* because *S2S+Att* aims to generate subject lines similar to the reference, whereas *BiSET* chooses a template beforehand and makes the generation fitting it, and thus the choice of the template may alter the style resulting in a non-desired styling.

Recommendation Results Since the goal of TemPEST is to improve the clicking rate of users, we evaluate the recommender system with TemPEST and the following baselines.

- *CDAE* (Wu et al. 2016) is short for collaborative denoising autoencoder, which uses denoising autoencoder to learn latent representation from user-item feedback.
- *GATE* is the gated attentive autoencoder (Ma et al. 2019) with source articles input.

Table 6: The user study of the proposed TemPEST and *BiSET*. The styles A to C indicate the style choices of users.

Models	User-based	BiSET
Style A	204	111
Style B	474	291
Style C	68	67
Total	746 (61.4%)	469 (38.6%)

Table 5 compares TemPEST with baselines in terms of $Recall@10$ and $NDCG@10$. The performance of TemPEST is better than CDAE, which demonstrates that the item content truly improves the recommendation results. Moreover, TemPEST outperforms GATE, which indicates that the user-specific article representation generated by TemPEST better captures the feature of the item description than the original embedding encoder of GATE.

Quantitative Results

User Study To evaluate the performance of subject generation in a real case, i.e., subjects conditioned on users preference, we conduct a user survey with 81 users. For the human evaluation, we consider the attractiveness. Given a pool of subjects, we ask users to choose the ones that he/she wants to click. Afterward, based on their choices of sequences, we generate subjects by TemPEST and compare with the ones generated by BiSET. The users are asked to choose one attracts him more. The results in Table 6 show that 61.4% of users think that the subjects generated by TemPEST are more attractive, while only 38.6% of users think that the headlines generated by BiSET (Wang, Quan, and Wang 2019) are more attractive. The chi-square significant test shows that under 90% of confidence, users significantly prefer the user-based sentences over the baseline despite the style they choose.

Case Study Table 7 shows a case study of the persoanlized subject generation task. Given a source article, the proposed TemPEST generate a user-specific subject according to the clicking history. BiSET tends to generate the subject that is similar to the real subject. As we take user-specific templates into consideration, the last three rows in Table 7 are different according to different user styles. For example, given user 1's style sequence, the results become more exaggerating like the word "Must-See" appearing in the sentence. While the outcome of user 2 also contains exaggerating words but includes the novel name in title. On the contrary, the result of user 3 tends to be a smoother one and it is also most similar to ground truth subject among three sentences. More details of dataset and case studies are shown in the supplementary material.[5]

Conclusions and Future Work

In this paper, we tackle the challenging task of generating an attractive subject with user-specific form and recommending

[5]https://github.com/yhchen2/TemPEST

Table 7: Example of the user-specific subject generated by our model. The Chinese descriptions on top are the real inputs of this dataset and translated to the English descriptions.

Article	《Wicked 綠野仙蹤女巫前傳》是紐約百老匯最賣座的音樂劇之一！精彩的故事改編自小說《女巫前傳》，加上實力派的演員及充滿設計感的舞台，是到紐約體驗百老匯音樂劇的不二選擇行程特色．《Wicked 綠野仙蹤女巫前傳》，紐約時報盛讚的劃時代精彩音樂劇．獲得多項殊榮，包括3項百老匯東尼獎、6項百老匯 Drama Desk 獎．國際級音樂劇演員搭配華麗的舞台效果，帶給觀眾多層次的震撼．免現場排隊買票，出示電子票券即可兌換實體門票 (*Wicked* is one of the most best-selling musical in Broadway theatre! Adapting from the novel *Wicked: The Life and Times of the Wicked Witch of the West*, with professional actors, the fashionable stage, and the wonderful story, it is the best choice if you want to experience Broadway musicals in New York. *Wicked*, an excellent musical highly praised by The New York Times, had received many awards, including 3 Tony Awards and 6 Drama Desk Awards. The performance acting by hotshot professional musical actors and the splendid stage effect would amaze the audience. Now you don't have to line up for the tickets, the electronic ticket can be exchanged to the real ticket directly.)
Real Subject	【百老匯經典音樂劇】Wicked綠野仙蹤女巫前傳門票 (Classic Broadway Musical: Tickets for *Wicked*)
BiSET	【紐約百老匯音樂劇】綠野仙蹤女巫前傳Wicked門票 (Broadway Musical in New York: Tickets for *Wicked*)
User 1	【紐約必去】綠野仙蹤女巫前傳門票 (New York Must-See: Tickets for *Wicked*)
User 2	【紐約必去景點】女巫前傳門票 (Must-Visit Attractions in New York: Tickets for *Wicked The Life and Times of the Wicked Witch of the West*)
User 3	【紐約經典音樂劇】綠野仙蹤女巫前傳門票 (Classic Musical in New York : Tickets for *Wicked*)

at the same time. To consider user information into personalized summarization, we propose a soft Template-based Personalized EDM Subject Generator (TemPEST). Extensive experiments on the real dataset show that TemPEST outperforms state-of-the-art methods. In the future, we plan to extend our model to cope with out-of-vocabulary problems.

Acknowledgement

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-108-2221-E-009-088, MOST-1108-2622-E-009-026-CC2, MOST-108-2634-F-009-006 and MOST-108-2218-E-009-050, and by the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan through grant 108W267.

References

Cao, Z.; Li, W.; Li, S.; and Wei, F. 2018a. Retrieve, rerank and rewrite: soft template based neural summarization. In *ACL*.

Cao, Z.; Wei, F.; Li, W.; and Li, S. 2018b. Faithful to the original: fact aware neural abstractive summarization. In *AAAI*.

Chopra, S.; Auli, M.; and Rush, A. M. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL-HLT*.

Gao, S.; Chen, X.; Li, P.; Ren, Z.; Bing, L.; Zhao, D.; and Yan, R. 2019. Abstractive text summarization by incorporating reader comments. In *AAAI*.

Jadhav, A., and Rajan, V. 2018. Extractive summarization with swap-net: sentences and words from alternating pointer networks. In *ACL*.

Kishore Papineni, Salim Roukos, T. W., and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; and Rush, A. M. 2017. OpenNMT: open-source toolkit for neural machine translation. In *ACL*.

Li, X., and She, J. 2017. Collaborative variational autoencoder for recommender systems. In *KDD*.

Li, J.; Li, H.; and Zong, C. 2019. Towards personalized review summarization via user-aware sequence network. In *AAAI*.

Lin, C.-Y., and Och, F. J. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING*.

Lin, C.-Y. 2004. Rouge: a package for automatic evaluation of summaries. In *Text summarization branches out*.

Liu, T.; Li, H.; Zhu, J.; Zhang, J.; and Zong, C. 2018. Review headline generation with user embedding. In *CCL*. Springer.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Ma, C.; Kang, P.; Wu, B.; Wang, Q.; and Liu, X. 2019. Gated attentive-autoencoder for content-aware recommendation. In *WSDM*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. In *ICLR*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Nallapati, R.; Zhou, B.; dos Santos, C.; Gulcehre, C.; and Xiang, B. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *SIGNLL*.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL-HLT*.

Ni, J., and McAuley, J. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *ACL*.

Sproat, R., and Emerson, T. 2003. The first international chinese word segmentation bakeoff. In *SIGHAN*.

Truong, Q.-T., and Lauw, H. 2019. Multimodal review generation for recommender systems. In *WWW*.

Vo, T. V., and Soh, H. 2018. Generation meets recommendation: proposing novel items for groups of users. In *RecSys*.

Vulić, I., and Moens, M.-F. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *SIGIR*.

Wang, K.; Quan, X.; and Wang, R. 2019. Biset: Bi-directional selective encoding with template for abstractive summarization. In *ACL*.

Wang, H.; Wang, N.; and Yeung, D.-Y. 2015. Collaborative deep learning for recommender systems. In *KDD*.

Wu, Y.; DuBois, C.; Zheng, A. X.; and Ester, M. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In *WSDM*.

Zhou, L., and Hovy, E. 2004. Template-filtered headline summarization. In *Text summarization branches out*.

Zhou, Q.; Yang, N.; Wei, F.; and Zhou, M. 2017. Selective sncoding for abstractive sentence summarization. In *ACL*.