PENS: A Dataset and Generic Framework for Personalized News Headline Generation

Xiang Ao[♠♢][*], Xiting Wang[♡], Ling Luo[♠♢], Ying Qiao[♣], Qing He[♠♢], Xing Xie[♡][†]
[♠]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[♡]Microsoft Research Asia
[♣]Microsoft
[♢]University of Chinese Academy of Sciences, Beijing 100049, China
{aoxiang,luoling18s,heqing}@ict.ac.cn
{xitwan,yiqia,xing.xie}@microsoft.com

Abstract

In this paper, we formulate the personalized news headline generation problem whose goal is to output a user-specific title based on both a user's reading interests and a candidate news body to be exposed to her. To build up a benchmark for this problem, we publicize a large-scale dataset named PENS (PErsonalized News headlineS). The training set is collected from user impressions logs of Microsoft News, and the test set is manually created by hundreds of native speakers to enable a fair testbed for evaluating models in an offline mode. We propose a generic framework as a preparatory solution to our problem. At its heart, user preference is learned by leveraging the user behavioral data, and three kinds of user preference injections are proposed to personalize a text generator and establish personalized headlines. We investigate our dataset by implementing several state-of-the-art user modeling methods in our framework to demonstrate a benchmark score for the proposed dataset. The dataset is available at https://msnews.github.io/pens.html.

1 Introduction

News headline generation (Dorr et al., 2003; Lopyrev, 2015; Alfonseca et al., 2013; Tan et al., 2017; See et al., 2017; Zhang et al., 2018; Xu et al., 2019; Murao et al., 2019; Gavrilov et al., 2019; Gu et al., 2020; Song et al., 2020), conventionally considered as a paradigm of challenging text summarization task, has been extensively explored for decades. Their intuitive intention is to empower the model to output a condensed generalization, e.g., one sentence, of a news article.

The recent year escalation of online content vendors such as Google News, TopBuzz, and etc (LaRocque, 2003) propels a new research direction that how to decorate the headline as an irresistible invitation to users for reading through the article (Xu et al., 2019) since more readings may acquaint more revenue of these platforms. To this end, specified stylized headline generation techniques were proposed, such as question headline (Zhang et al., 2018), sensational headline (Xu et al., 2019) generation, and so on (Shu et al., 2018; Gu et al., 2020). However, the over-decorate headlines might bring negative effects as click-baits begin to become notorious in ubiquitous online services[1].

Hence, the question is now changing to how to construct a title that catches on reader curiosity without entering into click-bait territory. Inspired by the tremendous success of personalized news recommendation (An et al., 2019; Wang et al., 2018; Li et al., 2010; Zheng et al., 2018) where the ultimate goal is to learn users' reading interests and deliver the right news to them, a plausible solution to this question could be producing headlines satisfying the personalized interests of readers.

It thus motivates the study of the personalized news headline generation whose goal is to output a user-specific title based on both a user's reading interests and a candidate news body to be exposed to her. Analogous to personalized news recommendations, user preference can be learned by leveraging the behavioral data of readers on content vendors, and the representation could personalize text generators and establish distinct headlines, even with the same news body, for different readers.

However, it might be difficult to evaluate the approaches of personalized headline generation due to the lack of large-scale available datasets. First, there are few available benchmarks that simultaneously contain user behavior and news content to train models. For example, most available news rec-

[*] This work was done when Xiang was visiting MSRA supported by the MSRA Young Visiting Researcher Program.

[†] Corresponding author.

[1] https://www.vizion.com/blog/do-clickbait-titles-still-work/

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 82–92
August 1–6, 2021. ©2021 Association for Computational Linguistics

ommendation datasets may predominately contain user-side interaction data, e.g., exposure impressions and click behaviors, but the textual features usually have already been overly pre-processed (Li et al., 2010; Zheng et al., 2018). As a result, advanced NLP techniques that extract useful features from textual data are limited. News headline generation datasets, on the other hand, usually consist of news bodies as well as their headlines, which all come from the news-side (Tan et al., 2017; Zhang et al., 2018) rather than the user-side. Though the MIND dataset (Wu et al., 2020), which was presented by Microsoft, simultaneously contains the user-side behavioral data and the news-side original textual data, it was constructed for personalized news recommendations rather than our problem. The more challenging issue for evaluating personalized headline generation approaches is the severe cost during the test phase. It could be intractable and infeasible to do an A/B test for every model in online environments. An efficient and fair testbed to evaluate the models in an offline mode is in urgent demand to make the effectiveness and reproducibility of proposed models comparable.

To this end, we publicize a dataset named PENS (PErsonalized News headlineS) in this paper as a benchmark to testify the performance of personalized news headline generation approaches. The training set of PENS is collected from the user impression logs of Microsoft News[2], in which $500,000$ impressions over $445,765$ users on more than one hundred thousand English news articles are provided. In addition, we collected 103 English native speakers' click behaviors as well as their more than $20,000$ manually-crafted personalized headlines of news articles on the same news corpus for testing. These manually-written headlines are regarded as the gold standard of the user-preferred titles. Then, proposed methods can take prevailing matching metrics, e.g., ROUGE, BLEU and etc., to verify the performance.

Moreover, we propose a generic framework to inject personalized interests into a proposed neural headline generator to enable a beacon for this area, considering there are few existing works that can generate personalized news headlines. In more detail, we devise three kinds of incorporation methods to inject user interest representation into a proposed neural headline generator with a transformer-based encoder and a pointer network-based (See et al., 2017) decoder. We implement six state-of-the-arts personalized news recommendation approaches to model user preferences and provide a horizontal standard for the PENS dataset. The experimental results show effective personalization modeling and comprehensive injection of user interests can underpin an improvement in the quality of personalized news headline generation. We expect PENS can serve as a benchmark for personalized headline generation and bolster the research in this area.

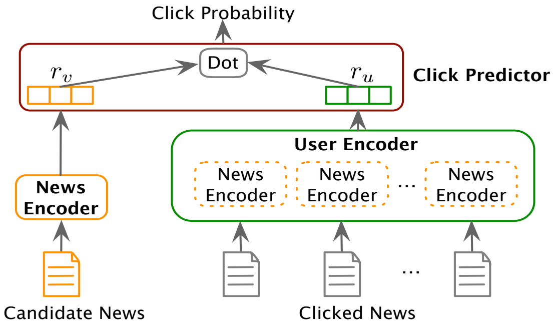

Figure 1: Personalized news recommendation framework.

2 Problem Formulation and Discussion

In this section, we formulate the problem of personalized news headline generation and differentiate it from personalized news recommendations.

2.1 Problem Formulation

The problem of personalized news headline generation is formulated as follows. Given a user u on an online content vendor, we denote his past click history as $[c_1^u, c_2^u, \ldots, c_N^u]$ where each c represents the headline of user u's clicked news and each headline is composed of a sequence of words $c = [w_{c_1}, \ldots, w_{c_T}]$ with the maximum length of T. Then, given the news body of a piece of news $v = [w_{v_1}, \ldots, w_{v_n}]$ to be exposed to user u, our problem is to generate a personalized news headline $H_v^u = [y_{v_1}^u, \ldots, y_{v_T}^u]$ based on the clicked news $[c_1^u, c_2^u, \ldots, c_N^u]$ and v.

2.2 Difference to Personalized News Recommendation

Here we differentiate our problem from personalized news recommendation whose general framework is shown as Fig. 1.

Recall that the aim of personalized news recommendation is computing and matching between the candidate news and the user's interests. Hence,

[2] https://microsoftnews.msn.com

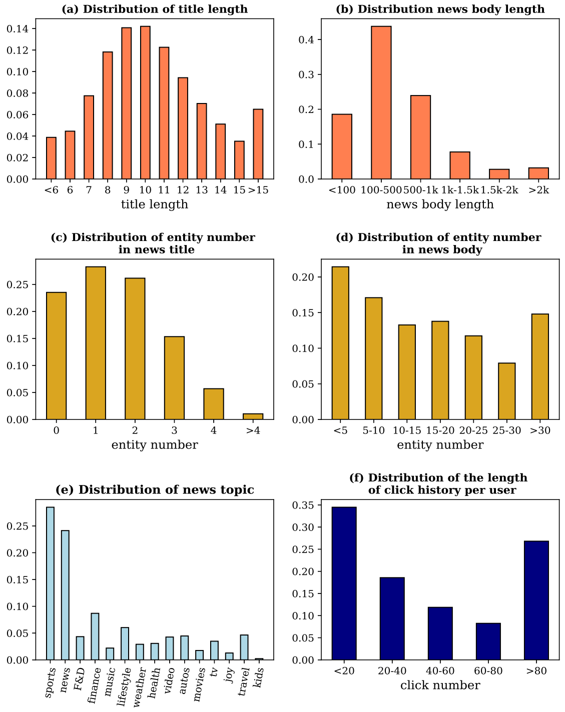

Figure 2: The statistics of news corpus and training set of the PENS dataset.

learning accurate news and user representations is critical for this problem. Under the neural framework, the news representation is usually modeled by a news encoder that encodes news title, news body or other attributes via various neural structures (Okura et al., 2017; Wang et al., 2018; Wu et al., 2019a; An et al., 2019; Wu et al., 2019a). The user representation is generated by engraving the high-level aspects over their clicked news sequences using sequential (Okura et al., 2017; An et al., 2019) or attentive modules (Wu et al., 2019b,a), in which every news is encoded by the news encoder in advance. Finally, the two representations are matched by the click predictor, and the whole model is trained by the supervision of click signals.

Different from personalized news recommendations, our personalized news headline generation could be regarded as an NLP task than a user modeling and matching problem. Although it similarly needs to model preferences for the individual users as what personalized news recommendations do, the output of our problem is a natural language sequence that the target user might be interested in, i.e., user-preferred news title, rather than a click probability score.

3 PENS Dataset

In this section, we detail our PENS dataset. The dataset was randomly sampled impression logs of Microsoft News from June 14 to July 12, 2019. Both user behaviors and news contents are involved, and each user was de-linked from the production system when securely hashed into an anonymous ID to reserve the data privacy issues.

3.1 News Corpus

The PENS dataset contains 113, 762 pieces of news articles whose topics are distributed into 15 categories. The topical distribution is demonstrated in Fig. 2 (c). Each news article in the PENS dataset includes a news ID, a title, a body and a category label. The average length of news title and news body is 10.5 and 549.0, individually. Moreover, we extract entities from each news title and body and link them to the entities in WikiData[3]. It could be taken as an auxiliary source to facilitate knowledge-aware personalization modeling and headline generation. The key statistical information of the PENS dataset is exhibited in Fig. 2 (a)–(e).

3.2 Training Set

The training set of PENS consists of impression logs. An impression log records the news articles displayed to a user as well as the click behaviors on these news articles when he/she visits the news website homepage at a specific time. We follow the MIND dataset (Wu et al., 2020) that we add the news click histories of every individual user to his/her impression log to offer labeled samples for learning user preferences. Hence, the format of each labeled sample in our training set is $[uID, tmp, clkNews, uclkNews, clkedHis]$, where uID indicates the anonymous ID of a user, tmp denotes the timestamp of this impression record. $clkNews$ and $uclkNews$ are the clicked news and un-clicked news in this impression, respectively. $clkedHis$ represents the news articles previously clicked by this user. All the samples in $clkNews$, $uclkNews$ and $clkedHis$ are news IDs, and they all sort by the user's click time. The histogram of the number of news in the clicked history per user is shown in Fig. 2 (f).

3.3 Test Set

To provide an offline testbed, we invited 103 English native speakers (all are college students) man-

[3] https://www.wikidata.org/wiki/Wikidata:MainPage

Table 1: The statistics of the training and test set in PENS. “wd.” in the table means word.

	#impression	#news	#user
Train	500,000	111,762	445,765
Test	NA	60,000	103
	avg. click/user	avg. wd./title	avg. wd./body
Train	74.8	10.5	549.0
Test	107.6	10.8	548.2

ually create a test set by two stages. At the first stage, each person browses $1,000$ news headlines and marks at least 50 pieces he/she is interested in. These exhibited news headlines were randomly selected from our news corpus and were arranged by their first exposure time. At the second stage, everyone is asked to write down their preferred headlines for another 200 news articles from our corpus, without exhibiting them the original news titles. Note that these news articles are excluded from the first stage, and only news bodies were exhibited to these annotators in this stage. These news articles are evenly sampled, and we redundantly assign them to make sure each news is exhibited to four people on average. The quality of these manually-written headlines was checked by professional editors from the perspective of the factual aspect of the media frame (Wagner and Gruszczynski, 2016). Low-quality headlines, e.g. containing wrong factual information, inconsistent with the news body, too-short or overlong, etc., are removed. The rest are regarded as the personalized reading focuses of these annotators on the articles and are taken as gold-standard headlines in our dataset. The statistics of the training and test sets of the PENS are shown in Table 1.

4 Our Framework

In this section, we illustrate our generic framework for resolving personalized news headline generation, and its key issue is how to inject the user preference into a news headline generator. We devise a headline generator with a transformer encoder and a pointer network decoder as our base model and propose three kinds of manners of injecting the user interests to generate personalized headlines. The user interests can be derived following the approaches in news recommendations community, and we omit its details due to the space limitation. The architecture of our proposed framework is shown as Figure 3.

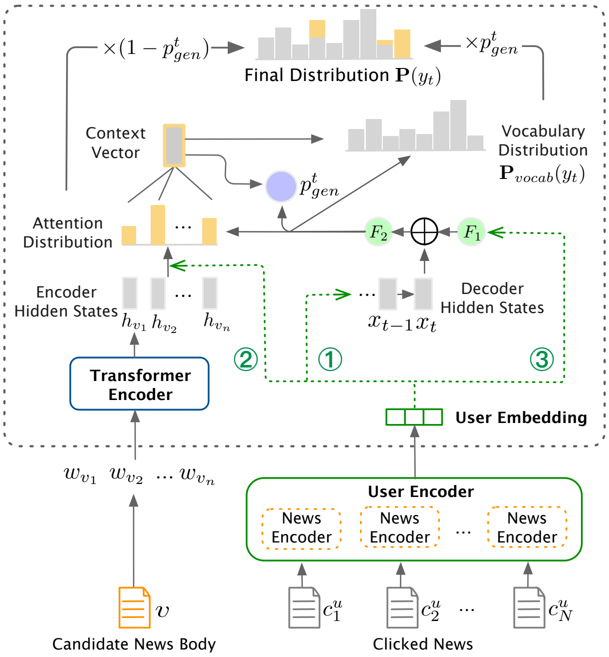

Figure 3: The generic framework of personalized news headline generation. Three kinds of user embedding injections are devised. ①: utilizing user embedding to initialize the decoder’s hidden state of the headline generator. ②: personalizing the attentive values on words in the news body by the user embedding. ③: perturbing the choice between generation and copying via the user embedding.

4.1 Headline Generator

The pin-point of our proposed headline generator is a variant of transformer encoder and pointer network decoder. During the encoding, given the news body of a candidate news $v = [w_{v_1}, \dots, w_{v_n}]$, its word embeddings $[e_{v_1}, \dots, e_{v_n}] \in R^{d_w}$ are first fed to a two-layer positional encoder. The first layer aims to enhance the word structure within the whole news body sequence following Vaswani et al. (2017), and we add the positional encoding to each embedding vector with,

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_w}) \tag{1}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_w}) \tag{2}$$

where pos is the word position and i is the dimension. We also apply a sentence-layer positional encoding to discover structural relations from higher level. Suppose the $W_{pos} \in R^{L \times d_s}$ represents the position embedding matrix of sentence level where L is the sentence length and d_s is the embedding size, the l-th row of W_{pos} represents the positional embedding of all the words in the l-th sentence. Thus, each word embedding e'_{pos} with positional information can be represented as:

$$e'_{pos} = (e_{pos} + PE_{pos}) \oplus W_{pos}[l]. \tag{3}$$

where $\oplus$ means concatenation. Furthermore, multi-head self-attention mechanism (Vaswani et al., 2017) is adopted to capture the word and sentence interactions by,

$$h_i = \text{softmax}(\frac{E'W_i^Q(E'W_i^K)^\top}{\sqrt{d_k}})E'W_i^V \tag{4}$$

where $d_k = \frac{d_s+d_w}{k}$ and $i = 1, \ldots, k$ given k heads. $W_i^Q, W_i^K, W_i^V \in R^{(d_s+d_w)\times d_k}$. E' represents the word sequence embeddings in candidate news v. Thus, the encoder hidden states $h = h_1 \oplus h_2, \ldots, h_k$ can be derived.

During the process of decoding, the decoded hidden state s_t at time step t can be derived after given the input x_t, and an *attention distribution* a_t over the encoder hidden states h is calculated as,

$$a_t = \mathcal{F}_\theta(h, s_t) \tag{5}$$

$$\mathcal{F}_\theta(h, s_t) = \text{softmax}(V_{att}^\top \text{tanh}(W_h h + W_s s_t + b_{att})) \tag{6}$$

where $\mathcal{F}_\theta$ represents a function template parameterized by θ to combine the linear transformation of the encoder and the decoder states, i.e., h and s_t. Next, the context vector c_t, which can be seen as a fixed-size representation read from the news body at time step t, is computed by a weighted sum of the encoder hidden states over the attention distribution. Then the *vocabulary distribution* is produced by,

$$P_{vocab}(w_t) = \text{tanh}(V_p[s_t; c_t] + b_v), \tag{7}$$

where V_p and b_v are learnable parameters while $P_{vocab}(w_t)$ represents the probability distribution over all the words in the vocabulary to predict the word at time step t.

Inspired by pointer-generator network (See et al., 2017), which exhibits desirable performance on either dealing with out-of-vocabulary (OOV) words or improving the reproducing factual details with copy mechanism, we adopt a pointer p_{gen}^t at decoding step t as a soft switch to choose between generating a word from the vocabulary with a probability of $P_{vocab}(w_t)$ or copying a word from the news body sampling from the attention distribution a_t. Thus, the probability distribution over the extended vocabulary is computed by,

$$P(w_t) = p_{gen}^t P_{vocab}(w_t) + (1 - p_{gen}^t) \sum_{j:w_j=w_t} a_{t,j} \tag{8}$$

where $P_{vocab}(w_t)$ is zero when w_t is out of vocabulary while $\sum_{j:w_j=w_t} a_{t,j} = 0$ when the w_t is not in the news body. p_{gen}^t is calculated based on the context vector c_t, decoder state s_t and the decoder input x_t:

$$p_{gen}^t = \mathcal{T}_\theta(c_t, s_t, x_t), \tag{9}$$

where $\mathcal{T}_\theta$ is a function template as Eq. (6).

4.2 Personalization by Injecting User Interests

So far, the imperative issue is to personalize the headline generator by injecting the user's preference. Recall that we can obtain user embedding indicating user's reading interests based on his/her historical clicked news sequences, and we denote such representation as u. As the user embedding u is usually not aligned with the word embeddings, it remains challenges to incorporate the user interests to influence the headline generation with personalized information.

In our framework, based on our headline generator, we propose three different manners to inject user interests, considering different intuitions, and they are exhibited in Fig. 3. First, the most simple and intuitive choice is to utilize the user embedding u to initialize the decoder hidden state of the headline generator. Second, under the empirical assumption that users may attend on different paragraphs and words in news articles corresponding to their individual preference, we inject u to affect the attention distribution a_t in order to personalize the attentive values on the different words in the news body. That is, we modify Eq. (5) and derive $a_t = \mathcal{F}_\theta(h, s_t, u)$. Lastly, we incorporate the personalized information to perturb the choice between generating a word from vocabulary or copying a word from the news body, and derive $p_{gen}^t = \mathcal{T}_\theta(c_t, s_t, x_t, u)$. Compared with Eq. (9), u is taken as an auxiliary parameter, where $\mathcal{T}_\theta$ is also a function template as Eq. (6).

4.3 Training

In this subsection, we present the training process of our framework. The headline generation can be considered as a sequential decision-making process, hence we optimize a θ parametrized policy for the generator by maximizing the expected reward of generated headline $Y_{1:T}$:

$$E_{Y_{1:T}\sim G_\theta}[R(Y_{1:T})]. \tag{10}$$

For the generator, policy gradient methods are applied to maximize the objective function in Eq. (10), whose gradient can be derived as,

$$\begin{aligned}\nabla_\theta J(\theta) \simeq E_{y_t \sim G_\theta(y_t|Y_{1:t-1})} \\ [\nabla_\theta \log G_\theta(y_t|Y_{1:t-1}) \cdot R(Y_{1:t-1}, y_t)]\end{aligned} \tag{11}$$

where the reward R is estimated by the degree of personalization, fluency and factualness as we aim to generate a user-specific and coherent headline to cover the main theme of news articles and arouse personalized reading curiosity. The implemented rewards in our framework contain: (1) The personalization of the generated headline is measured by the dot product between the user embedding and the generated headline representation. Such a score might imply a matching degree of personalization. (2) The fluency of a generated headline is assessed by a language model. We adopt a two-layer LSTM pre-trained by maximizing the likelihood of news body and consider the probability estimation of a generated headline as the fluency reward. (3) We measure the degree of factual consistency and the coverage by calculating the mean of ROUGE (Lin, 2004)-1, -2 and -L F-scores between each sentence in the news body and the generated headline, and then take the average of the top 3 scores as the reward. We average all three rewards as the final signal. As all the above reward functions only produce an end reward after the whole headline is generated, we apply a Monte Carlo Tree search to estimate the intermediate rewards.

5 Experimental Evaluation

In this section, we investigate our proposed PENS dataset and conduct several comparisons to give benchmark scores of personalized headline generation on this dataset. In the following part, we will introduce the compared methods first, and then detail the experimental setup, and finally present the results and analysis.

5.1 Compared Methods

We mainly compare two groups of approaches. The first group consists of various user modeling methods, which are all SOTA neural-based news recommendation methods: (1) **EBNR** (Okura et al., 2017) learns user representations by aggregating their browsed news with GRU. (2) **DKN** (Wang et al., 2018) is a deep knowledge-aware network for news recommendation. (3) **NPA** (Wu et al., 2019b) proposes personalized attention module in both news and user encoder. (4) **NRMS** (Wu et al., 2019c) conducts neural news recommendation with multi-head self-attention. (5) **LSTUR** (An et al., 2019) models long- and shor-term user representations based on user ID embedding and sequential encoding, individually. (6) **NAML** (Wu et al., 2019a) proposes multi-view learning in user representation.

To the best of our knowledge, there are no exclusive methods for personalized news headline generation. Hence we take several headline generation methods for comparison. (1) **Pointer-Gen** (See et al., 2017) proposes an explicit probabilistic switch to choose between copying from source text and generating word from vocabulary. (2) **PG+RL-ROUGE** (Xu et al., 2019) extends Pointer-Gen with as a reinforcement learning framework which generates sensational headlines by considering ROUGE-L score as rewards.

5.2 Experiment Setup

We perform the following preprocessings. For each impression, we empirically keep at most 50 clicked news to learn user preferences, and set the length of news headline and news body to 30 and 500, respectively. Word embeddings are 300-dimension and initialized by the Glove (Pennington et al., 2014) while the size of position embeddings at sentence level is 100. The multi-head attention networks have 8 heads.

First of all, we conduct news recommendation tasks to pretrain a user encoder with a learning rate of 10^{-4} on the first three weeks, i.e., from June 14 to July 4, 2019, on the training set, and test on the rest. Notice that the parameters of the user encoder are not updated thereafter. Meanwhile, the headline generator is also pretrained with a learning rate of 0.001 by maximizing the likelihood of original headlines based on a random but fixed user embedding which can be considered as a global user without personalized information. Next, we train each individual model for 2 epochs following Eq. 10, and Adam (Kingma and Ba, 2014) is used for model optimization where we sample 16 sequences for Monte Carlo search.

5.3 Evaluation Metrics

For news recommendation evaluation, we report the average results in terms of AUC, MRR, nDCG@5 and nDCG@10. For personalized headline generation, we evaluate the generation quality using F1 ROUGE (Lin, 2004) [4] including unigram

[4] We compute all ROUGE scores with parameters "-a -c 95 -m -n 4 -w 1.2." Refer to

Table 2: The overall performance of compared methods. "R-1, -2, -L" indicate F scores of ROUGE-1, -2, and -L, and "NA" denotes "Not Available". "IM" means injection methods, c.f. ①, ②, and ③ in Fig. 3 for details.

Methods	**Metrics**							
	AUC	**MRR**	**NDCG@5**	**NDCG@10**	**IM**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Pointer-Gen	NA	NA	NA	NA	NA	19.86	7.76	18.83
PG+RL-ROUGE	NA	NA	NA	NA	NA	20.56	8.42	20.03
EBNR	63.97	22.52	26.45	32.81	①	25.13	9.03	20.73
					②	25.49	9.14	20.82
					③	24.62	8.95	20.40
DKN	65.25	24.07	26.97	34.24	①	25.97	9.23	20.92
					②	27.48	10.07	21.81
					③	25.02	8.98	20.34
NPA	64.91	23.65	26.72	33.96	①	25.49	9.14	20.82
					②	26.11	9.58	21.40
					③	26.35	9.71	21.82
NRMS	64.27	23.28	26.60	33.58	①	24.92	9.01	20.75
					②	26.15	9.37	21.03
					③	25.41	9.12	20.91
LSTUR	62.49	22.69	24.71	32.28	①	23.71	8.73	21.13
					②	24.10	8.82	20.73
					③	23.11	8.42	20.38
NAML	66.18	25.51	27.56	35.17	①	27.49	10.14	21.62
					②	28.01	10.72	22.24
					③	27.25	10.01	21.40

and bigram overlap (ROUGE-1 and ROUGE-2) to assess informativeness, and the longest common subsequence (ROUGE-L) to measure fluency. Here we adopt ROUGE because we care more about evaluating the recall of the generated results. All the reported values are the averaged results of 10 independently repeated runs.

5.4 Experimental Results

Since we include six kinds of user modeling methods from personalized news recommendations and propose three ways of injecting user interests in our framework, we can derive 18 variants of approaches that can generate personalized news headlines. Meanwhile, there are two headline generation baselines, hence we totally have 20 methods for evaluation. The overall performance is illustrated in Table 2, and we have the following observations.

First, we can see that every personalized news headline generation method can outperform non-personalized methods like PG. It might be that our proposed framework can generate personalized news headlines by incorporating user interests. Such personalized headlines are more similar to the manually-written ones, which are taken as gold-standard in our evaluation. Second, we find that user modeling makes a difference in generating personalized headlines. For instance, NAML achieves the best performance in news recommendation by learning news and user representations from multiple views, i.e., obtaining 66.18, 25.51, 27.56 and 35.17 on AUC, MRR NDCG@5 and NDCG@10. Then injecting the user preferences learned by NAML to the proposed headline generator also gets the highest ROUGE scores with either way of the incorporation. We conjecture it is because better user modeling methods can learn more rich personalized information from click behaviors, and well-learned user embeddings could strive to generate better-personalized headlines. Third, it is reported that the second way of injecting user interests gets the best performance on most of the user modeling methods, e.g., EBNR, DKN and NAML. It is probably because the differentiation of the attention distribution is intensified after the user embedding perturbation, which then impacts the word generation in the decoding process. However, it still remains a large room for explorations on better injecting user representations into the generation process since the second way seems to be defective at some time.

https://pypi.python.org/pypi/pyrouge/0.1.3

Table 3: A case study on personalized headline generation for two different users by personalized (NAML+HG) and non-personalized (Pointer-Gen). Underlined words and colored words represent the correlated words in the manually-written headlines, clicked news, and the generated headlines, respectively.

Case 1. Original Headline:	Venezuelans rush to Peru before new requirements take effect
Pointer-Gen:	Venezuelans rush to Peru
user A written headline:	New **requirements** set to take effect causes **Venezuelans** to rush to **Peru**
NAML+HG for user A:	**Peru** has stricter entry **requirements** for escaping **Venezuelans** on that influx.
Clicked News of user A:	1. Peru and Venezuela fans react after match ends in a draw
	2. Uruguay v. Peru, Copa America and Gold Cup, Game threads and how to watch
user B written headline:	**Venezuelan migrants** to Peru face danger and discrimination
NAML+HG for user B:	Stricter entry requirements on **Venezuelan migrants** and **refugees**.
Clicked News of user B:	1. Countries Accepting The Most Refugees (And Where They're Coming From)
	2. Venezuelan mothers, children in tow, rush to migrate

5.5 Case Study

To further comprehend our task and the proposed framework, we demonstrate interesting cases from two representative methods, namely one non-personalized method Pointer-Gen (PG) and one personalized method NAML+HG which utilizes the second user interests injection (c.f. Fig. 3). We also exhibit the manually-written headlines by the users and the original news headline as references.

From the results shown in Table 3, we can observe that generated headline by non-personalized method might omit some detailed but important information. We believe the reason is that PG is trained via supervised learning to maximize the log-likelihood of ground-truth news headlines. While our framework is trained via RL technique where coverage score is considered as an indicator to encourage the generation to be more complete. In addition, the exhibited cases show that our framework can produce user-specific news headlines in accordance with their individual interests reflected by historical click behaviors. Meanwhile, some key phrases in the personalized-written titles successfully appeared in the machine-generated headlines.

6 Related Work

Headline generation has been considered as specialized text summarization (Luo et al., 2019; Jia et al., 2020), from which both extractive (Dorr et al., 2003; Alfonseca et al., 2013) and abstractive summarization (Sun et al., 2015; Takase et al., 2016; Tan et al., 2017; Gavrilov et al., 2019; See et al., 2017) approaches prevailed for decades. Extractive methods select a subset of actual sentences in original article, which may derive incoherent summary (Alfonseca et al., 2013). While abstractive models, basically falling in an encoder-decoder (Shen et al., 2017a; Murao et al., 2019) framework, can generate more condensed output based on the latent representation of news content. However, the nature of text summarization methods without considering interactions between news and users renders them ineffective in our personalized headline generation.

Recently, stylized headlines generation were proposed to output eye-catching headlines by implicit style transfer (Shen et al., 2017b; Fu et al., 2018; Prabhumoye et al., 2018) or style-oriented supervisions (Shu et al., 2018; Zhang et al., 2018; Xu et al., 2019). However, either training a unified text style transfer model or constructing a personalized text style transfer model for every user is infeasible due to the complex personalized style-related patterns and the limited personalized-oriented examples. Meanwhile, these methods might suffer from the risk of entering into click-bait territory.

Personalized News Recommendation is also related to our problem. Among them, content-based recommendations (Okura et al., 2017; Liu et al., 2010; Li et al., 2011; Lian et al., 2018; Wang et al., 2018; Wu et al., 2019a,b) perform user and news matching on a learned hidden space, and user representation is learned based on historical clicked news contents. It inspires us to personalize headline generator by incorporating user embeddings. Deep models (Lian et al., 2018; Wang et al., 2018; Wu et al., 2019b,a), recently, demonstrated significant improvements because of their capabilities in representation learning on both user-side and news-side data. Different from the efforts on personalized news recommendation, our work focuses on generating fascinating headlines for different users, which is orthogonal to existing work.

7 Conclusion and Future Work

In this paper, we formulated the problem of personalized news headline generation. To provide an offline testbed for this problem, we constructed a dataset named PENS from Microsoft News. The news corpus of this dataset contains more than 100 thousand news articles over 15 topic categories. The training set constitutes of $500,000$ impressions of $445,765$ users to learn user interests and construct personalized news headline generator by distant supervisions. The test set was constructed by 103 annotators with their clicked behaviors and manually-written personalized news headlines. We propose a generic framework that injects user interests into an encoder-decoder headline generator in three different manners to resolve our problem. We compared both SOTA user modeling and headline generating approaches to present benchmark scores on the proposed dataset.

For future work, we first believe designing more complex and refined approaches to generated more diversified personalized news headlines will be interesting. More importantly, how to improve personalization while keeping factualness will be another interesting work, and it will propel the methods deployable in practical scenarios. Third, news headline personalization might burgeon the news content personalization, which is a more challenging but interesting open problem.

Acknowledgments

The research work is supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002104, the National Natural Science Foundation of China under Grant No. 92046003, 61976204, U1811461. Xiang Ao is also supported by the Project of Youth Innovation Promotion Association CAS and Beijing Nova Program Z201100006820062.

References

Enrique Alfonseca, Daniele Pighin, and Guillermo Garrido. 2013. HEADY: News headline abstraction through event pattern clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1243–1253, Sofia, Bulgaria. Association for Computational Linguistics.

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long- and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345, Florence, Italy. Association for Computational Linguistics.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 on Text Summarization Workshop - Volume 5*, HLT-NAACL-DUC '03, page 1–8, USA. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*, pages 663–670.

Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive model for headline generation. In *ECIR*, pages 87–93.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, You Wu, Cong Yu, Daniel Finnie, Hongkun Yu, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating representative headlines for news stories. In *WWW*, pages 1773–1784.

Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3622–3631, Online. Association for Computational Linguistics.

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

Paul LaRocque. 2003. *Heads you win: An easy guide to better headline and caption writing*. Marion Street Press, Inc.

Lei Li, Dingding Wang, Tao Li, Daniel Knox, and Balaji Padmanabhan. 2011. Scene: a scalable two-stage personalized news recommendation system. In *SIGIR*, pages 125–134.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670.

Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2018. Towards better representation learning for personalized news recommendation: a multi-channel deep fusion approach. In *IJCAI*, pages 3805–3811.

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

Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *IUI*, pages 31–40.

Konstantin Lopyrev. 2015. Generating news headlines with recurrent neural networks. *arXiv preprint arXiv:1512.01712*.

Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like HER: Human reading inspired extractive 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 3033–3043, Hong Kong, China. Association for Computational Linguistics.

Kazuma Murao, Ken Kobayashi, Hayato Kobayashi, Taichi Yatsuka, Takeshi Masuyama, Tatsuru Higurashi, and Yoshimune Tabuchi. 2019. A case study on neural headline generation for editing support. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 73–82, Minneapolis, Minnesota. Association for Computational Linguistics.

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *KDD*, pages 1933–1942.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Shi-Qi Shen, Yan-Kai Lin, Cun-Chao Tu, Yu Zhao, Zhi-Yuan Liu, Mao-Song Sun, et al. 2017a. Recent advances on neural headline generation. *Journal of Computer Science and Technology*, 32(4):768–784.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017b. Style transfer from non-parallel text by cross-alignment. In *NIPS*, pages 6830–6841.

Kai Shu, Suhang Wang, Thai Le, Dongwon Lee, and Huan Liu. 2018. Deep headline generation for clickbait detection. In *ICDM*, pages 467–476.

Yun-Zhu Song, Hong-Han Shuai, Sung-Lin Yeh, Yi-Lun Wu, Lun-Wei Ku, and Wen-Chih Peng. 2020. Attractive or faithful? popularity-reinforced learning for inspired headline generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8910–8917.

Rui Sun, Yue Zhang, Meishan Zhang, and Donghong Ji. 2015. Event-driven headline generation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Beijing, China. Association for Computational Linguistics.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on Abstract Meaning Representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1054–1059, Austin, Texas. Association for Computational Linguistics.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *IJCAI*, pages 4109–4115.

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.

Michael W Wagner and Mike Gruszczynski. 2016. When framing matters: How partisan and journalistic frames affect individual opinions and party identification. *Journalism & Communication Monographs*, 18(1):5–48.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *WWW*, pages 1835–1844.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. In *IJCAI*, pages 3863–3869.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019b. Npa: Neural news recommendation with personalized attention. In *KDD*, pages 2576–2584.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019c. Neural news recommendation with multi-head self-attention. 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 6389–6394, Hong Kong, China. Association for Computational Linguistics.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020.

MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics.

Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? sensational headline generation with auto-tuned reinforcement learning. 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 3065–3075, Hong Kong, China. Association for Computational Linguistics.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Huanhuan Cao, and Xueqi Cheng. 2018. Question headline generation for news articles. In *CIKM*, pages 617–626.

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. Drn: A deep reinforcement learning framework for news recommendation. In *WWW*, pages 167–176.