Generating User-Engaging News Headlines

Pengshan Cai,[1*] **Kaiqiang Song,**[2] **Sangwoo Cho,**[2] **Hongwei Wang,**[2]
Xiaoyang Wang,[2] **Hong Yu,**[1,3] **Fei Liu,**[4] **Dong Yu**[2]

[1]University of Massachusetts, Amherst [2]Tencent AI Lab, Bellevue, WA
[3]University of Massachusetts, Lowell [4]Emory University
{pengshancai,hongyu}@cs.umass.edu fei.liu@emory.edu
{riversong,swcho,hongweiw,shawnxywang,dyu}@global.tencent.com

Abstract

The potential choices for news article headlines are enormous, and finding the right balance between conveying the essential message and capturing the reader's attention is key to effective headlining. However, presenting the same news headline to all readers is a suboptimal strategy, because it does not take into account the different preferences and interests of diverse readers, who may be confused about why a particular article has been recommended to them and do not see a clear connection between their interests and the recommended article. In this paper, we present a novel framework that addresses these challenges by incorporating user profiling to generate personalized headlines, and a combination of automated and human evaluation methods to determine user preference for personalized headlines. Our framework utilizes a learnable relevance function to assign personalized signature phrases to users based on their reading histories, which are then used to personalize headline generation. Through extensive evaluation, we demonstrate the effectiveness of our proposed framework in generating personalized headlines that meet the needs of a diverse audience. Our framework has the potential to improve the efficacy of news recommendations and facilitate creation of personalized content.[1]

1 Introduction

Personalized news recommendation systems, such as Google News and Yahoo News, help users discover articles that align with their interests (Karimi et al., 2018). However, these systems often present the same article headline to all users, making it difficult for them to understand the connection between their interests and the recommended article, potentially reducing the effectiveness of the recommendation system. To address this, we propose a new framework for generating *personalized, engaging headlines* that clearly show the connection between a user's reading history and a recommended article. Our framework has the potential to improve the efficacy of personalized news recommendations, and recommendations for short videos, articles, recipes, *etc.* (Majumder et al., 2019; Kanouchi et al., 2020; Gosangi et al., 2021)

Generating personalized headlines is a challenging task due to the constraints of conciseness and the need to capture the reader's attention. A personalized headline should (a) effectively convey the main message of the article and (b) provide a clear link to the user's reading history, using only about 10 words on average (Bernstein et al., 2020). There are two main challenges in this task. First, a headline that entices users to click, but only presents limited information and fails to convey the essential story, becomes clickbait rather than a useful headline (Bourgonje et al., 2017; Potthast et al., 2018). Second, it is difficult to find large scale annotated datasets containing news articles, multiple personalized headlines, and associated user profiles. Such a dataset would be useful in developing personalized headlines, but it is currently unattainable.

The key to effective personalization is to develop a *comprehensive framework* that enables us to (a) understand users' interests based on their reading histories, (b) produce personalized headlines, and (c) evaluate the effectiveness of these headlines in terms of user preference. Previous studies on headline generation have primarily focused on producing headlines that accurately summarize a given news article or its first sentence (Song et al., 2018; Xu et al., 2019; Matsumaru et al., 2020; Song et al., 2021; Kanungo et al., 2021), but have not considered the potential benefits of personalization. In this study, we propose a pipeline that incorporates user profiling[2] and a comprehensive synthesis of

*Work completed during an internship at Tencent AI Lab

[1]Our code can be accessed publicly at: https://github.com/pengshancai/user-engaging-headlines.

[2]We are interested in analyzing users' reading histories, i.e., the sequence of news headlines they have recently browsed, to gain a deeper understanding of their interests and preferences. We do not have access to users' demographic data.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3265–3280

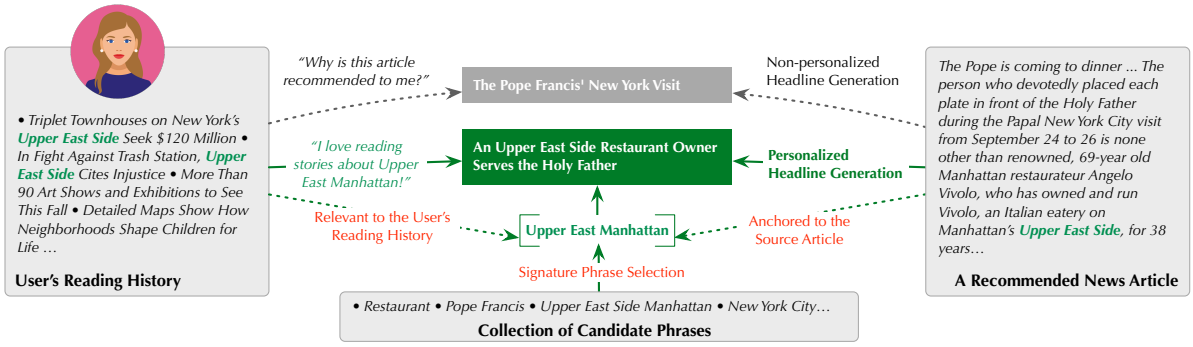

Figure 1: An example of generating a personalized news headline using our framework (black solid line) as compared to generating general headlines directly from the news article (grey dotted line). Both headlines are appropriate for the news article, but headline 1 is more attractive to users interested in the topic *Upper East Side, Manhattan*.

automated and human evaluation methods for user preference to produce personalized headlines that cater to a varied audience.

Our approach focuses on learning a relevance function that condenses a user's reading history into a collection of signature phrases. This method for user profiling is both efficient and adaptable, as the signature phrases can be easily updated as the user's interests evolve (Bansal et al., 2015). These signature phrases are derived from news article based on the user's reading history through contrastive learning *without the need for annotated data.* For example, if the phrase *Upper East Side* frequently appears in the user's reading history, it could become a signature phrase for that user (Figure 1). These signature phrases *do not need to appear verbatim* in the user's reading history and can indicate broader interests, e.g., if the phrases *Avengers* and *Hulk* appear in the user's reading history, it could indicate a love for Marvel movies and *Marvel Studios* could be a signature phrase that reflects this interest. We build a synthetic dataset that trains the model to generate personalized headlines for a news article. Using signature phrases, our model is able to create a connection between the recommended article and the user's interests, resulting in personalized headlines that are both engaging and anchored to the article to avoid clickbait.

Evaluating personalized news headlines presents unique challenges (Gligorić et al., 2021). It would be ideal to have human evaluators judge the effectiveness of system headlines. Indeed, we have conducted a human evaluation in this study. However, this process is time-consuming and costly, making it impractical during the system development phase. Thus, we propose *a comprehensive synthesis of automated and human evaluation methods* to assess headline relevance and user preference. By using signature phrases, we can synthesize user profiles of various types. We hypothesize that personalized headlines generated for these user profiles will be preferred by the same users over generic, non-personalized headlines according to recommender-driven metrics (Karpukhin et al., 2020; Wu et al., 2021a). We also experiment with a variety of automatic metrics to assess headline quality in terms of informativeness, relevance to the source article, and content accuracy (Kryscinski et al., 2020; Fabbri et al., 2021).

In this paper, we make the following contributions:

- we present a comprehensive framework for generating personalized news headlines that convey the essential message of the article and capture the reader's attention while also aligning with their interests. Our framework utilizes a learnable relevance function to derive signature phrases from users' reading histories and uses them to personalize the headlines;
- we thoroughly synthesize automated and human evaluation methods to assess the effectiveness of headlines in terms of their accuracy and user preference. We further compare our proposed framework with strong headline generation baselines, present results on benchmark news datasets, and identify promising directions for future research through an in-depth analysis of system outputs.

2 Related Work

Automatic headline generation has made significant progress in recent years (Matsumaru et al., 2020; Horvitz et al., 2020; Laban et al., 2021; Song et al., 2020; Goyal et al., 2022), thanks in part to the development of large language models (Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020a; Brown et al., 2020; Chowdhery et al., 2022) and the availability of benchmark news datasets such as Gigaword, XSum, and Newsroom (Rush et al., 2015; Narayan et al., 2018; Grusky et al., 2018).

These datasets include a single headline for each news article, serving as the groundtruth for the models. In contrast to previous works, we aim to personalize headline generation to improve content recommendations, where a personalized headline should convey the main points of the article and capture the user's attention.

Personalization is a highly sought-after technique, and researchers have explored its use for tasks such as headline generation, dialog response generation and recipe creation (Ao et al., 2021; Majumder et al., 2019; Flek, 2020; Wu et al., 2021b; Dudy et al., 2021). We anticipate that this technique to continue to have a significant impact. For example, when a recommender system distributes news articles or short videos, personalizing the headline can help users find a clear connection between their interests and the recommended article/video (Karimi et al., 2018; Bernstein et al., 2020), thus improving their experience.

Evaluating personalized content is a largely under-explored area, partly due to the lack of ground truth for personalized content generation (Gligorić et al., 2021). Without ground truth, it is challenging to apply commonly used text generation evaluation metrics such as ROUGE, BLEU, BERTScore, MoverScore, BLEURT, etc. (Lin, 2004; Post, 2018; Zhang et al., 2020b; Zhao et al., 2019; Sellam et al., 2020). To leverage recent advances in data synthesis (Pasunuru et al., 2021; Amplayo and Lapata, 2020; Magooda and Litman, 2021), we propose synthesizing user profiles of various types. We then evaluate system headlines against these profiles along multiple dimensions, including their alignment with user interests, relevance to the source article, and content accuracy. In the following, we provide details of our approach.

3 Our Approach

Our goal is to generate a user-engaging headline that conveys the main idea of a given news article d for a specific user u. To achieve this, we have developed a three-step framework: (1) *Signature phrases identification*. Using a key-phrase generation module, we identify a set of candidate signature phrases $Z_d = \{z_1, z_2, \dots\}$ that cover various aspects of d (Section 3.1); (2) *User signature phrases selection*. From the set of candidate signature phrases, we select a subset $Z_d^u \subseteq Z_d$ that relates to user u's interests as the user signature phrases (Section 3.2); (3) *Signature-oriented headline generation*. Based on the news article d and the selected user signature phrases Z_d^u, we generate a headline that introduces the content of the article d from the perspective of the user u's personalized interests (Section 3.3).

3.1 Signature Phrases Identification

We approach this task as a conditional text generation problem, in which the model takes a news article or headline as input and outputs all candidate signature phrases in the input sequence, separated by semicolons. We use a BART model that has been pretrained on the KPTimes dataset[3]. KPTimes (Gallina et al., 2019) is a large-scale dataset containing 279K news articles paired with editor-curated signature phrases. Unlike other datasets for signature phrase identification (Meng et al., 2017; Krapivin et al., 2009) that focus on scientific research papers, KPTimes focuses on extracting signature phrases in news articles, making it well-suited for our task. The model is trained by minimizing the cross-entropy loss between the predicted signature phrase sequences and the human-curated signature phrase sequences.

3.2 User Signature Selection

In this step, we rank all candidate signature phrases in Z_d based on their level of engagement with user u's reading history H_u, and select the top k candidate signature phrases as the user signature phrases. Suppose that the user's history H_u can be defined as a set of headlines of articles that the user has previously read, i.e., $H_u = \{t_1, t_2, \dots\}$. We first convert each signature phrase $z_i \in Z_d$ into a dense vector $\mathbf{z}_i$ using a signature phrase encoder. To calculate the user-engaging scores for each candidate signature phrase z_i, we consider two different encoding strategies for the user's history:

(1) **Holistic history encoding**. We concatenate all headlines in the user's reading history H_u with additional semicolons for headline separation. Then we encode the concatenated headlines into a dense vector $\mathbf{h}_u$ using a holistic history encoder. The engaging score $S(z_i, H_u)$ of a signature phrase $z_i \in Z_d$ for user u is obtained by the dot product of the two vectors:

$$S(z_i, H_u) = \mathbf{z}_i^\top \mathbf{h}_u. \quad (1)$$

(2) **Individual history encoding**. Each individual headline $t_j \in H_u$ is encoded as a dense vector $\mathbf{t}_j$ using an individual headline encoder. The user-engaging score is then defined as the maximum dot-product relevance between the signature phrase z_i

[3] https://huggingface.co/ankur310794/bart-base-keyphrase-generation-kpTimes

and each individual headline in the reading history:

$$S(z_i, H_u) = \max_{t_j \in H_u} \mathbf{z}_i^\top \mathbf{t}_j. \quad (2)$$

In practice, we train the user signature phrase selection model using an in-batch contrastive learning approach (Radford et al., 2021). We consider a batch of synthesized users $\{u_1, u_2, \cdots, u_{N_B}\}$ where N_B is the batch size, and each user u_i has exactly one user signature phrase z_i. The reading history H_i for user u_i is then constructed by randomly sampling news articles whose candidate signature phrases contain z_i, i.e., $H_i = \{d \mid z_i \in Z_d\}$. In this way, (z_i, H_i) is considered as a positive pair, and (z_i, H_j) $(i \neq j)$ is considered as a negative pair. The contrastive loss for this batch is defined as follows:

$$L_{select} = \frac{1}{2}\left(\sum_{i=1}^{N_B} \log \frac{S(z_i, H_i)}{\sum_{j=1}^{N_B} S(z_i, H_j)} + \right. \quad (3)$$

$$\left. \sum_{j=1}^{N_B} \log \frac{S(z_j, H_j)}{\sum_{i=1}^{N_B} S(z_i, H_j)}\right) \quad (4)$$

3.3 Signature-Oriented Headline Generation

We model the user-specific headline generation process as a conditional generation task. Given a news article d and a user u, along with the user signature phrases $Z_d^u \subseteq Z_d$, our goal is to generate a headline $t = [w_1, w_2, \ldots]$ for d, where w_i is the i-th token in t. The loss for this generation step is calculated as the negative log-likelihood of the conditional language generation:

$$L_{gen} = -\sum_i \log \Pr(w_i \mid w_1, \cdots, w_{i-1}; Z_d^u, d) \quad (5)$$

Specifically, the input to the generator is the concatenation of the user signature phrases Z_d^u and news article d, and the output is the signature-based headline t. During the training stage, Z_d^u is identified from t, the ground-truth headline of d. During the inference stage, Z_d^u is identified from d itself and selected by user signature selection models, since the headline t is not available before generation. We use BART here as the generator for headline generation.

4 Corpora Processing

In this section, we describe the corpora processing step, including the creation of synthesized users and the generation of signature phrase based headlines. Our data is sourced from two existing news corpora: Newsroom (Grusky et al., 2018) and Gigaword (Rush et al., 2015; Graff et al., 2003). The Newsroom corpus contains 995,041 article-headline pairs in its training set, 108,837 in its validation set, and 108,862 in its test set. The Gigaword corpus contains 7,704,419 instances in its training set, 394,390 in its validation set, and 381,045 in its test set. For each corpus, we construct two datasets: a synthesized user dataset and a headline generation dataset. The first dataset is used for training the use signature phrase selection model (Section 3.2) and evaluating the entire system, while the second dataset is used for training the signature-oriented headline generation model (Section 3.3). Further data statistics can be found in Table 1.

Corpus		Newsroom	Gigaword
Synthesized user dataset			
Train	# instances	994,680	6,848,000
	# signature phrases per user	1	1
	Avg. # articles read by a user	16.17	16.31
Dev	# instances	49,860	49,984
	# signature phrases per user	1	1
	Avg. # articles read by a user	16.32	16.33
Test	# instances	10,000	10,000
	# signature phrases per user	1~5	1~5
	Avg. # articles read by a user	15.03	14.99
Headline generation dataset			
# train instances		995,041	7,704,419
# dev instances		58,530	394,390
Avg. # words/article		661.58	421.42
Avg. # words/headline		8.73	8.44
Avg. # signature phrase/article		11.36	10.81
Total # of signature phrases		48,820	25,084

Table 1: Statistics of the datasets. For each corpus, the synthesized user dataset is used for training the signature phrase selection module and evaluating the entire system, while the headline generation dataset is used for training the headline generation module (it does not have a test set because the generation step is evaluated in the entire system using the test set of synthesized user dataset).

Synthesized User Creation. As real user data is not available, we generate synthesized users to mimic real users' reading histories. The process for creating synthesized users is illustrated in Figure 2 and consists of the following steps: (1) Identification of signature phrases in all news articles of a corpus to build a candidate phrase pool; (2) Mapping of each signature phrase to a series of news articles that contain that phrase; (3) Random sampling of a subset of phrases from the candidate phrase pool as each synthesized user's area of interest; (4) Random sampling of a set of news articles that contain each user's chosen interest phrase using the phrase-article map established in step 2.

During the training stage of the signature phrase selector, each synthesized user is assigned only one

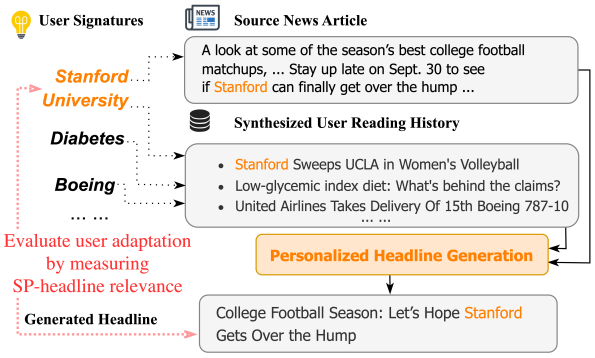

Figure 2: Synthesizing user profiles. The synthesized user's interests contain randomly selected interest phrases, i.e. *Stanford University*, *Diabetes*, *Boeing*. etc. Some news headlines related to these phrases are chosen to represent the synthesized user's reading history. During the inference stage, one news article containing the interest phrase *Stanford University* is selected as the source article for headline generation.

interest phrase to enable contrastive learning (Eq. 4). However, when evaluating the model, each synthesized user is assigned $1 \sim 5$ interest phrases to mimic real-world scenarios. It is important to note that it is easier to generate personalized headlines for users with simpler backgrounds (e.g. users whose reading histories only relate to one or two topics). To study the effect of the number of users' interested phrases on the generated headlines, we create 2,000 synthesized users with $1 \sim 5$ number of interested phrases respectively.

In general, headline personalizing is only effective when the source article content aligns with the user's interests. To ensure relevancy, we randomly select one of the user signature phrases from each synthesized user, and then randomly choose one news article that contains the selected phrase as the input for the test case. This ensures that the news article whose headline needs to be generated is relevant to the user. The evaluation details are further explained in Section 5.

Headline Generation. In order to generate signature phrase oriented headlines, we use the signature phrases identification model to extract signature phrases from the original headlines. These generated phrases, along with the corresponding news article contents, are then fed into the headline generation model to generate the original headlines. In our experiments, we truncate all news articles to a maximum of 512 tokens and only keep signature phrases that appear in more than 10 news articles. On average, around 10 candidate signature phrases are identified in each news article, providing a diverse range of perspectives for headline generation.

5 Experiments

We thoroughly evaluate our proposed system from different perspectives, including objective evaluation (Section 5.2), subjective evaluation (Section 5.3) and ablation studies (Section 5.4), for personalized headline generation.

5.1 Baseline Methods

We compare the performance of our system with the following baseline approaches: (1) *PENS-EBNR* and (2) *PENS-NRMS* (Ao et al., 2021) are LSTM-based personalized headline generation models. Both were trained on the PENS dataset, but using different reading history encoding models; (3) *Vanilla System* is a BART-large model fine-tuned directly on headline generation datasets without using signature phrases; (4) *Vanilla Human* refers to original headline given by the author of the news article; (5) *SP-headline* uses signature phrases identified in the original human-written headline to guide headline generation; (6) *SP-random* randomly selects signature phrases in the news article to guide headline generation. (7) *SP-holistic* and (8) *SP-individual* were introduced in previous sections.

5.2 Objective Evaluation

We use various metrics to evaluate the entire personalized headline generation pipeline:
(1) *Relevance Metrics*. We use pre-trained DPR (Karpukhin et al., 2020) and Sentence-BERT (Reimers and Gurevych, 2019) models to calculate the relevance score between texts. Specifically, we report dot-product similarity when using DPR, and cosine similarity when using Sentence-BERT. These relevance metrics are calculated for both the *headline-user relevance* and the *headline-article relevance*. For *headline-user relevance*, the score is calculated between the generated headline and the user signatures. For *headline-article relevance*, the score is calculated between the generated headline and the entire news article.
(2) *Recommendation Score*. Following (Wu et al., 2021a), we train a news recommendation system using the MIND dataset (Wu et al., 2020). The system takes in a user's reading history and a headline of a news article, and outputs a score indicating the degree to which the system would recommend the news to the user.
(3) *Factual Consistency*. We apply the pre-trained FactCC model (Kryscinski et al., 2020) to obtain the factual consistency score between the generated

Methods		User Adaptation Metrics			Article Loyalty Metrics			Other Metrics		
		H-U Relevance			H-A Relevance					
		DPR	SBERT	REC Score	DPR	SBERT	FactCC	R-L	Ext Cvrg	Length
Newsroom										
Baselines	PENS-NRMS	50.85	0.221	2.449	60.25	0.659	0.498	17.98	0.982	9.99
	PENS-EBNR	50.89	0.219	2.476	60.84	0.666	0.521	19.75	0.984	10.00
	Vanilla System	51.78	0.249	2.697	64.31	**0.681**	0.639	37.02	0.828	8.51
	Vanilla Human	51.39	0.241	2.690	64.00	0.642	0.682	N/A	0.749	8.96
Ours	SP Headline	52.42	0.270	2.577	63.74	0.651	0.694	42.63	0.772	7.53
	SP Random	52.26	0.263	2.735	64.31	0.652	0.680	29.40	0.817	8.87
	SP holistic-N	53.23	0.286	2.896	64.33	0.654	0.673	29.52	0.817	8.83
	SP individual-N	54.19	0.313	2.735	64.57	0.659	0.670	30.14	0.818	8.87
	SP holistic-F	54.00	0.310	2.882	64.24	0.655	0.662	29.92	0.814	8.79
	SP individual-F	**55.05**	**0.342**	**2.947**	**64.85**	0.658	**0.695**	29.83	0.820	8.98
Gigaword										
Baselines	PENS-NRMS	52.30	0.22	3.144	63.72	0.678	0.524	23.06	0.999	9.97
	PENS-EBNR	52.51	0.221	3.224	64.51	0.696	0.551	22.30	0.997	10.00
	Vanilla System	53.28	0.241	**3.526**	**66.90**	**0.702**	0.636	44.95	0.797	8.22
	Vanilla Human	52.80	0.236	3.489	66.08	0.652	0.684	N/A	0.716	8.57
Ours	SP Headline	52.94	0.236	3.478	66.39	0.684	0.655	54.68	0.782	8.13
	SP Random	52.44	0.235	3.216	64.33	0.625	0.718	33.33	0.764	7.86
	SP holistic-N	53.39	0.253	3.414	64.81	0.638	0.697	35.39	0.768	7.84
	SP individual-N	54.08	0.272	3.455	65.25	0.648	0.695	36.36	0.776	7.87
	SP holistic-F	54.14	0.278	3.396	64.77	0.636	0.704	35.16	0.769	7.87
	SP individual-F	**54.82**	**0.299**	3.459	65.34	0.643	**0.738**	34.65	0.778	8.06

Table 2: Objective evaluation results of all methods. "-F" means using the fine-tuned signature phrase encoder, headline encoder and user history encoder, while "-N" means using the naive DPR models as encoders. "REC Score" refers to recommendation score. Vanilla approaches do not consider human preference.

headline and the news article. We report the percentage of generated headlines that are predicted to be factually consistent with the news article by the FactCC model.

(4) *Surface Overlap*. We use ROUGE-L F1 and Extractive Coverage to evaluate the surface overlap between the generated headline and the reference headline/news article. ROUGE (Lin, 2004) scores are widely used to evaluate the surface level coverage of generated summaries against golden standards. Specifically, ROUGE-L F1 measures the longest common sub-sequence between the generated output and reference. Extractive Coverage (Grusky et al., 2018) is the percentage of words in the generated headline that are from the source news article, measuring the extent to which the summary is derived from the text.

Table 2 presents objective evaluation results for generated headlines. We elaborate our observations from the following perspectives:

User Adaptation. (1) The methods *SP holistic* and *SP individual* generally show better performance, indicating that our signature phrase based headline generation framework is able to generate more user-oriented headlines. In contrast, while *Vanilla System* and *SP Headline* achieve higher Rouge-L scores, they have lower scores in user adaptation, suggesting that they have higher similarity with the original headline but do not achieve personalization. (2) Comparing SP based methods, we observe that using selectors fine-tuned on our signature selection datasets (i.e. *-F*) leads to more user-preferred headlines than their naive counterparts (i.e. *-N*). This reflects the improvement of fine-tuning signature phrase selector. It is worth noting that the performance of *SP Random* is significantly lower than *SP holistic/individual*, and almost similar to *Vanilla System*, which suggests that user adaptation is only achieved when signature phrases of users' interests are well-selected. (3) *SP individual* shows better performance than *SP holistic*, indicating that individual encoding better aligns users' reading history with their interests.

Article Loyalty. (1) While *Vanilla System* generally achieves better performance in headline-article relevance, *SP individual-F* generates more headlines that are identified as factually consistent by FactCC. Our analysis found that headlines generated by our SP-based methods are usually anchored to news articles by the signature phrase, i.e. the generated headlines may contain content in the context of the signature phrase (as shown in the example in Figure 2). This keeps the generated headlines related and factually consistent with the news article, thus avoiding click-bait headlines. (2) The extractive converge of the original human headlines is lower than all machine-generated headlines, which implies that human written headlines are more abstractive. This explains the original headlines' low performance in article loyalty metrics. Note that ROUGE scores do measure our goal of headline personalization, we present the results only to show

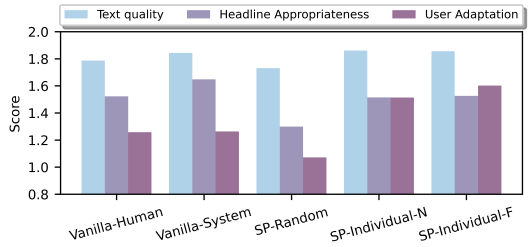

Figure 3: Result of human evaluation scores on the generated headlines w.r.t. text quality, headline appropriateness, and user adaptation.

the generated headlines' surface-level resemblance to the human written ones.

5.3 Subjective Evaluation

We conduct a two-step human evaluation using 16 evaluators who have high English proficiency. In the first step, we collected 2,260 news headlines from 113 common topics in Newsroom and Gigaword corpus. We presented the volunteers with the article headlines and corresponding topics and asked them to select around 20 headlines of their interests mimicking their interest phrases and reading histories. In the second step, we generated headlines for 12 randomly selected news articles containing the volunteers' interested phrases (6 from Newsroom and 6 from Gigaword). We then asked the volunteers to evaluate the generated headlines through the following five approaches: (1) *Vanilla Human*; (2) *Vanilla System*; (3) *SP-random*; (4) *SP-individual-N*; (5) *SP-individual-F*. We evaluated the headlines from three perspectives: (1) *User adaptation*; (2) *Headline appropriateness* and (3) *Text quality*. The grading scale ranges from 1 (worst) to 3 (best), and detailed grading standards are provided in Appendix A.3.

According to Figure 3, our signature-oriented headline generation approaches, *SP-Individual-F* and *SP-Individual-N*, perform better than other baseline methods in terms of user adaptation. This is in line with the objective results that our signature-oriented framework generates headlines that cater more to users' interests.

Further, the headlines generated by *Vanilla System* obtain the highest scores in headline appropriateness. However, after analyzing the generated headlines, we realized that some identified signature phrases did not correlate well with the article's main point, thus diverging from the article. For example, in the third example in Table 3, the generated headline focuses on *Shanghai Index's drop*, which is only a minor evidence to support the article's main point, i.e. *China's stock market crush*, and is therefore not appropriate to be included in the headline.

Moreover, the *Vanilla Human* did not receive the highest scores. We found some of the human written headlines are overly rhetorical and not easily understandable to ordinary readers (see the fourth example in Table 3). All NLP models achieve good performance (around 1.8 points) in text quality, which is similar to the scores of the human-written headlines. [4]

1	**User Signatures**: Mark Zuckerberg; Bill Gates **News Article**: The Giving Pledge, invented by Bill and Melinda Gates and Warren Buffett to spur the philanthropy of billionaires, ... assuredly the coolest recruits are Facebook co-founders Mark Zuckerberg and Dustin Moskovitz, who each turned 27 in May ... **Generated Headline**: The Giving Pledge: Zuckerberg and Gates at 27
2	**User Signatures**: The Force Awakens **User Interest Phrase**: Star Wars **News Article**: Star Wars: Episode 7 has revealed its full title - it will be called Star Wars: The Force Awakens ... **Generated Headline**: Star Wars Episode 7 to be called Star Wars: The Force Awakens
3	**User Signatures**: Shanghai Composite Index **News Article**: China stocks fell more than 1 percent on Tuesday morning ... the Shanghai Composite Index lost 1.4 percent ... **Generated Headline**: Shanghai Composite Index falls 1.4% despite market-soothing measures
4	**User Signatures**: Photography **News Article**: ... Self-publishing is not a new development in photography, but recently the trend to make, edit, design and produce ... **Human Headline**: Self-publish or be damned: why photographers are going it alone **Generated Headline**: Self-published photography books to be showcased at Photographers' Gallery

Table 3: Examples of generated headlines.

Selector	Hit@1	Hit@3	Hit@5	Mean Rank↓
Newsroom				
Random	9.28	27.79	46.28	5.071
Holistic-N	18.30	41.82	57.95	4.395
Holistic-F	30.10	54.69	68.81	3.376
Individual-N	30.99	57.05	71.68	3.193
Individual-F	**40.34**	**67.57**	**79.64**	**2.395**
Gigaword				
Random	9.28	27.79	46.28	5.071
Holistic-N	16.91	39.56	58.31	4.142
Holistic-F	29.21	55.44	70.95	3.094
Individual-N	23.98	50.09	67.50	3.438
Individual-F	**34.05**	**64.01**	**79.71**	**2.426**

Table 4: The impact of different signature phrase selectors.

Moreover, the *Vanilla Human* did not receive the highest scores. We found some of the human written headlines are overly rhetorical and not easily understandable to ordinary readers (see the fourth example in Table 3). All NLP models achieve good performance (around 1.8 points) in text quality, which is similar to the scores of the human-written headlines. [4]

5.4 Ablation Study

Selectors Evaluation. To evaluate the performance of signature selection, we rank all candidate signature phrases within an article for a synthesized user and report the following metrics: (1) Hit@K, which is the percentage of times that the correct signature phrase is ranked among the top K; (2) Mean rank, which is the average rank of the correct signature phrase. We use our synthesized user evaluation dataset to evaluate both headline generation and signature selection.

[4] We present more examples in Appendix A.4.

# User's Interest Phrases	User Adaptation Metrics			Article Loyalty Metrics			Other Metrics		
	H-U Relevance		REC Score	H-A Relevance		FactCC	R-L	Ext Cvrg	Length
	DPR	SBERT		DPR	SBERT				
1	55.63	0.362	4.532	65.14	0.665	70.2	30.28	0.826	9.04
2	55.04	0.347	3.077	64.87	0.656	69.2	30.03	0.818	9.02
3	54.96	0.343	2.555	64.84	0.660	68.5	29.55	0.821	9.04
4	54.96	0.330	2.262	64.53	0.653	68.9	29.31	0.815	8.82
5	54.65	0.328	2.310	64.88	0.658	70.7	29.97	0.821	8.98
10	54.39	0.323	1.871	64.96	0.655	69.3	29.18	0.813	8.89
20	53.74	0.305	1.65	64.7	0.657	66.9	30.01	0.812	8.93
30	53.14	0.291	1.778	64.66	0.658	69.1	29.55	0.817	8.94

Table 5: Result of generated headlines for newsroom articles when synthesized users have different number of interest phrases.

Methods	User Adaptation Metrics			Article Loyalty Metrics			Other Metrics		
	H-U Relevance		REC Score	H-A Relevance		FactCC	R-L	Ext Cvrg	Length
	DPR	SBERT		DPR	SBERT				
History Oriented (GPT-3)	51.76	0.277	4.277	64.05	0.676	0.64	29.99	0.751	7.02
Topic Oriented (GPT-3)	52.73	0.296	4.562	64.21	**0.685**	0.65	26.32	0.759	7.80
SP individual-F	**54.75**	**0.330**	**4.618**	**64.85**	0.672	**0.71**	36.89	0.835	9.14

Table 6: Performance of GPT-3 generated headlines compared to our *SP individual-F*.

History Oriented: Assume a reader has already read a series of articles titled [Title 1], [Title 2], ... Here's an input news article: [Article]. Generate a compelling headline within ten words for this news article that the reader would find interesting.

Topic Oriented: [Article]. Generate a compelling headline within ten words for the above news article that a reader who has already read a series of articles on the topics of [Topic 1], [Topic 2], ... would find interesting.

Table 7: Two paradigms of applying GPT-3 in personalized headline generation. *History Oriented* uses GPT-3 to generate headlines for users based on their reading history. *Topic Oriented* first obtains focused signature phrases using our signature identification and selection modules, and then generates the headline based based on the focused topics using GPT-3.

As shown in Table 4, *Individual-F* demonstrates the best performance among all selectors. This explains the high user adaptation scores of headlines generated by *SP individual-F*. We have observed that the selector does not always choose the gold user signature phrases, yet the generated headline still relates to user's interests. For example, in the second example of Table 3, even though the user's interested phrase *Star War* was not chosen as the user signature, the generated headline is still relevant to *Star War*, as the selected signature phrase *The Force Awakens* is the subheading of a movie in the *Star War* movie series.

Factors Affecting Headline Generation. Through our experiments, we have identified that the following factors affect the quality of the generated headlines: (1) Number of topics that the user is interested in. As shown in Table 5[5], the evaluation results of headlines generated from newsroom articles for synthesized users with varying number of interest phrases indicates that, as the number of interest phrases increases, the user adaptation scores decreases, while other scores remain roughly the same. This suggests that it is easier to generate personalized headlines for users who read news related to fewer interest phrases. However, even when the number of interest topics increases to 30, our proposed method still achieves better user adaptation scores then the vanilla systems, while showing similar performance in article loyalty metric. (2) Number of user signature phrases. Our analysis of generated headlines revealed that when the signature-oriented headline generator takes multiple user signature phrases as input, the generated headline may contain factual errors. This is because the generator is compelled to incorporate irrelevant signature phrases into a coherent headline, as seen in the first example in Table 3). As a result, we only use a single signature phrase to guide headline generation.

Applying GPT-3 for Personalized Headline Generation. Recently, GPT-3 (Brown et al., 2020) has been found to be effective in zero-shot prompting automatic summarization (Goyal et al., 2022). In this section, we investigate whether prompts can inspire GPT-3[6] to generate personalized headlines of good quality. To achieve this goal, we conduct experiment with 100 random samples from our newsroom test set using two paradigms, as shown in Table 7, and present the results in Table 6.

Our *SP individual-F* method outperforms GPT-3 based methods in terms of user adaptation metrics and ROUGE-L score. This suggests that despite GPT-3's strong ability in zero-shot setting, it is still

[5]In this experiment, we additionally include 3 groups of synthesized users who has 10/20/30 interest topics, each single user has 50-60 news in their reading histories.

[6]In our experiment, we use OpenAI's text-davinci-003.

incomparable to models that are specifically trained for our headline generation task. Specifically, the *topic oriented* method shows better performance in user adaptation metrics than the *history oriented* method, which implies that our topic selector effectively reveals users' interests.

6 Conclusion

We investigate the generation of personalized headlines tailored to various users' interests. We propose a topic-focused generation framework and methods for creating synthesized data to support the training of our framework without the need for human-annotated datasets. Additionally, we explore evaluation methods that enable the automatic evaluation of the generated headlines from multiple perspectives. Our experiments demonstrate the effectiveness of our proposed approaches.

7 Limitations

Personalized news headline generation has the potential to improve the way users consume and understand the news. However, it is important to be aware of its limitations. The performance of any natural language generation model, including those used for personalized news headlines, is dependent on the quality and consistency of the data used to train it. Similar to personalized recommendation systems, personalized headlines have the potential to create echo chambers. If the model is trained on a biased or unrepresentative dataset, it may generate outputs that are incomplete, inaccurate, or misleading. Therefore, it is crucial to be aware of the limitations of the model and to ensure that it is trained on high-quality data to generate accurate and personalized headlines.

8 Ethical Considerations

It is important to use the proposed personalized news headline generation technique ethically and responsibly. While the technique aims to improve personalized content recommendations and optimize the user experience, it could also be used to generate headlines that are more likely to appeal to an individual reader, potentially resulting in a biased view of the news. In this paper, we have taken necessary precautions to protect personal data. Our technique is based on a user's reading history, which is represented as a sequence of recently viewed news headlines. No demographic data such as age, gender, or location is used or collected, due to privacy concerns. We encourage the community to continue to explore the potential risks and implications of this technique.

References

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online. Association for Computational Linguistics.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

Trapit Bansal, Mrinal Das, and Chiranjib Bhattacharyya. 2015. Content driven user profiling for comment-worthy recommendations of news and blog articles. In *Proceedings of the 9th ACM Conference on Recommender Systems*, page 195–202.

Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, and Katharina A Zweig. 2020. Diversity, fairness, and data-driven personalization in (news) recommender system. *Dagstuhl perspectives workshop 19482*.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89, Copenhagen, Denmark. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, and Sebastian Gehrmann et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Ygor Gallina, Florian Boudin, and Beatrice Daille. 2019. KPTimes: A large-scale dataset for keyphrase generation on news documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 130–135, Tokyo, Japan. Association for Computational Linguistics.

Kristina Gligorić, George Lifchits, Robert West, and Ashton Anderson. 2021. Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report Protocol). *PLoS One*, 16(9):e0257091.

Rakesh Gosangi, Ravneet Arora, Mohsen Gheisarieha, Debanjan Mahata, and Haimin Zhang. 2021. On the use of context for predicting citation worthiness of sentences in scholarly articles. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4539–4545, Online. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Zachary Horvitz, Nam Do, and Michael L. Littman. 2020. Context-driven satirical news generation. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 40–50, Online. Association for Computational Linguistics.

Shin Kanouchi, Masato Neishi, Yuta Hayashibe, Hiroki Ouchi, and Naoaki Okazaki. 2020. You may like this hotel because ...: Identifying evidence for explainable recommendations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 890–899, Suzhou, China. Association for Computational Linguistics.

Yashal Shakti Kanungo, Sumit Negi, and Aruna Rajan. 2021. Ad headline generation using self-critical masked language model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 263–271, Online. Association for Computational Linguistics.

Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems – survey and roads ahead. *Information Processing Management*, 54(6):1203–1227.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Mikalai Krapivin, Aliaksandr Autaeu, and Maurizio Marchese. 2009. Large dataset for keyphrases extraction.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Lucas Bandarkar, and Marti A. Hearst. 2021. News headline grouping as a challenging NLU task. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3186–3198, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ahmed Magooda and Diane Litman. 2021. Mitigating data scarceness through data synthesis, augmentation and curriculum for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2043–2052, Punta

Cana, Dominican Republic. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2019. Generating personalized recipes from historical user preferences. 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 5976–5982, Hong Kong, China. Association for Computational Linguistics.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346, Online. Association for Computational Linguistics.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Vancouver, Canada. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. 2021. Data augmentation for abstractive query-focused multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13666–13674.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. 2018. Crowdsourcing a large corpus of clickbait on Twitter. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1498–1507, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kaiqiang Song, Bingqing Wang, Zhe Feng, and Fei Liu. 2021. A new approach to overgenerating and scoring abstractive summaries. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1392–1404, Online. Association for Computational Linguistics.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Liu Ren, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

Kaiqiang Song, Lin Zhao, and Fei Liu. 2018. Structure-infused copy mechanisms for abstractive summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1717–1729, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021a. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1652–1656.

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.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021b. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, 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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

Signature Phrase Selector. We fine-tune pre-trained DPR models on our signature phrase selection datasets (both Newsroom and Gigaword) to obtain signature phrase selectors. The pre-trained models were obtained from huggingface. Under individual setting, the signature phrase encoder was initialized from the DPR question encoder[7], and the headline encoder was initialized from the DPR context encoder [8]. (The DPR models were also applied in evaluating headline-user & headline-article relevance.) Our signature selectors and headline generators are trained on 8 Nvidia-A100 GPUs. Under holistic setting, the signature phrase encoder was initialized from the DPR question encoder, and the history encoder was initialized from the DPR context encoder. Fine-tuning key hyper-parameters are shown in Table 8:

Signature Phrase Selection	
Batch size	96 * 8
Learning rate	3e-5
# of train epochs	15
Signature phrase max length	16 tokens
Headline max length	48 tokens
Reading history max length	256 tokens
Signature-oriented Headline Generation	
Batch size	48 * 8
Learning rate	5e-5
# of train epochs	6
Input news article max length	512 tokens
Reading history max length	256 tokens

Table 8: Hyperparameters of the model.

Signature-oriented Headline Generator. We fine-tune a pre-trained BART-large model[9] on our user-oriented headline generation dataset. Our key hyper-parameters are shown in Table 8.

PENS. The PENS baselines were implemented following the original paper's github repo [10]. For comparison fairness, we only use the headline of each news article to represent that article in the user's reading history. We limited the max length of the generated headlines to be 10 words. Other then than that we train the models following the repo's original setting.

Sentence BERT. We use the pre-trained sentence BERT model (all-MiniLM-L6-v2) from the following repo: https://github.com/UKPLab/sentence-transformers The original sentence BERT setting is to calculate the semantic similarity between two sentences. As a result, when calculating the headline-article relevance, we report the maximum similarity score between the headline and all sentences in the news article.

Recommender System. As no pretrained model was provided by the authors We train the model from scratch. We use the implementation provided by https://github.com/wuch15/PLM4NewsRec with default settings.

FactCC. The FactCC model we apply as an evaluation metric was obtained from the following paper's original github repo (directly use the pre-trained model): https://github.com/salesforce/factCC.

GPT-3. We apply GPT-3 by calling OpenAI API

[7] https://huggingface.co/facebook/dpr-question_encoder-single-nq-base

[8] https://huggingface.co/facebook/dpr-ctx_encoder-single-nq-base

[9] https://huggingface.co/facebook/bart-large

[10] https://github.com/LLluoling/PENS-Personalized-News-Headline-Generation

at https://openai.com/api/.

A.2 Analysis of GPT-3 Generated Headlines

In addition to the findings we reported in section 5.4, we report the following observations of headlines generated by GPT-3 guided by prompts: We found including the phrase *within ten words* in the prompt greatly boost the quality of the generated headlines. When including this phrase, the average length of the generated headlines is less than 8 words. However, when not including this phrase, the average length of generated headlines is close to 15 words, which is much longer than the average length of human written news headlines (around 8 words). Long headlines can contain too much information, and does not fulfill the headline requirement of being succinct.

A.3 Human Evaluation Details

We explain human evaluation criteria in Table 10.

A.4 A Case Study

Table 9 shows examples of editor-written, generic headlines compared to headlines generated by our proposed system.

Example 1 shows the smartphone market rankings can be approached from different perspectives. The editor headline focuses on Apple's slip to 3rd place, while the generated headline emphasizes on Xiaomi's rise to the top. In this case, the generated headline aligns better with the reader's interests.

In Example 2, both the human headline and generated headline mention Sony's new PC. Our generated headline includes a reference to Microsoft, making it likely to capture the reader's interest.

In Example 3, we show that the generated headline has a stronger correlation with the news content compared to the human-written headline.

	Example 1
News Article	Apple has hit a road bump in it quest to dominate the Chinese smartphone market, according to data tracking the shipment of phones in the second quarter. Over the period from April to June, Fortune's leading startup unicorn Xiaomi regained its label as the largest smartphone vendor in China by capturing a 15.9% market share, ... Right behind was Huawei with a 15.7% share ...
Human Headline	Apple Slips To 3rd Place In Key China Smartphone Market
Generated Headline	Xiaomi reclaims top spot in China smartphone market (Signature phrase: Xiaomi)
	Example 2
News Article	Thin and light is in, and nobody is pushing that more than Sony this holiday season. On Tuesday morning, the company announced the pricing and availability for what just may be the most intriguing item in its holiday lineup, the Tap 11 tablet PC ... It's perhaps the jewel of Sony's holiday lineup, and it just might be able to go head-to-head with Microsoft's Surface 2 thanks to that ultra-light profile and the inclusion of the keyboard cover...
Human Headline	Sony announces Tap 11 tablet PC, Flip laptop lines
Generated Headline	Sony unveils lightest tablet PC yet, taking on Microsoft's Surface 2 (Signature phrase: Microsoft)
	Example 3
News Article	Luxury resorts from Thailand to Germany to California are offering a range of detox fasting programmes aimed at weight loss and well-being, but the "health" factor remains open to question. Shunning food for religious or spiritual reasons has existed for centuries, as during Ramadan, Lent or Yom Kippur for instance ...
Human Headline	To eat or not to eat
Generated Headline	Dieting holidays: 'detoxification' or 'health' fad? (Signature phrase: Diet)
	Example 4
News Article	A study of New York City's pioneering law on posting calories in restaurant chains suggests that when it comes to deciding what to order, people's stomachs are more powerful than their brains ... It found that about half the customers noticed the calorie counts, which were prominently posted on menu boards ... But when the researchers checked receipts afterward, they found that people had, in fact, ordered slightly more calories than the typical customer had before the labeling law went into effect, in July 2008.
Human Headline	Calorie Postings Don't Change Habits, Study Finds
Generated Headline	Calories on Menu Boards May Not Cut Obesity, Study Finds (Signature phrase: Obesity)
	Example 5
News Article	It's a loaded question, one with no clear answer. But in the year since Apple's co-founder and visionary CEO died, it's been asked in tech circles over and over: Who is the next Steve Jobs? ... Bezos actually has a host of traits that mirror Jobs. Like Jobs was with Apple, he's the founder of Amazon as well as its CEO ...
Human Headline	Who is the next Steve Jobs (and is there one)?
Generated Headline	Amazon's Bezos: The next Steve Jobs? (Signature phrase: Jeff Bezos)

Table 9: Human written headlines vs. generated headlines.

User Adaptation: Does the headline cater to the user's interest	
2	The headline is related to user's interest
1	The headline is weakly related to user's interest
0	The headline is not related to user's interest at all
Headline Appropriateness: Is the headline proper to the news article	
2	The headline is proper to the news article
1	The headline is not entirely appropriate
0	The headline does not correlate to the news article at all
Text quality: Is the headline grammatically and semantically correct	
2	The headline has no semantic or grammar error
1	The headline has one minor semantic or grammar error
0	The headline has serious semantic or grammar errors

Table 10: Each summary is scored on a scale of 0 (worst) to 2 (best) for three criteria: relevance to the user, appropriateness of the headline, and overall text quality.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☑ A2. Did you discuss any potential risks of your work?
8

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

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

B ☑ Did you use or create scientific artifacts?

3, 4

☑ B1. Did you cite the creators of artifacts you used?
1, 2, 3, 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

5

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

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

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

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

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5, Appendix

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

5, Appendix

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
I attached it in the supplementary material (data.zip)

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
The authors recruit their friends as volunteer evaluators

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
We explain to evaluators that their personal data will not be disclosed

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The risk and potential consequences of exposing personal information is low

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
5