Hooks in the Headline: Learning to Generate Headlines with Controlled Styles

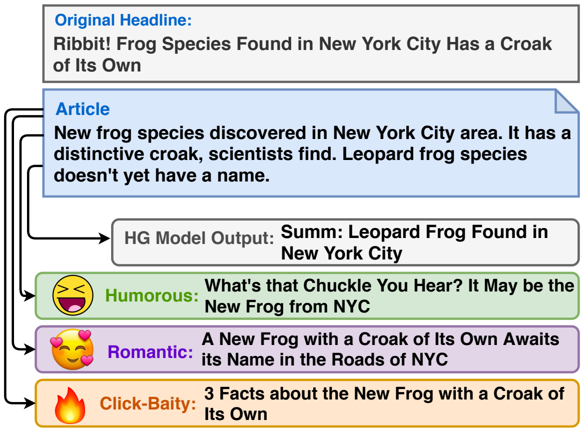
Di Jin,[1] Zhijing Jin,[2] Joey Tianyi Zhou,[3*] Lisa Orii,[4] Peter Szolovits[1]
[1]CSAIL, MIT, [2]Amazon Web Services, [3]A*STAR, Singapore, [4]Wellesley College
{jindi15,psz}@mit.edu, zhijing.jin@connect.hku.hk
zhouty@ihpc.a-star.edu.sg, lorii@wellesley.edu

Abstract

Current summarization systems only produce plain, factual headlines, but do not meet the practical needs of creating memorable titles to increase exposure. We propose a new task, Stylistic Headline Generation (SHG), to enrich the headlines with three style options (humor, romance and clickbait), in order to attract more readers. With *no* style-specific article-headline pair (only a standard headline summarization dataset and mono-style corpora), our method TitleStylist generates style-specific headlines by combining the summarization and reconstruction tasks into a multitasking framework. We also introduced a novel parameter sharing scheme to further disentangle the style from the text. Through both automatic and human evaluation, we demonstrate that TitleStylist can generate relevant, fluent headlines with three target styles: humor, romance, and clickbait. The attraction score of our model generated headlines surpasses that of the state-of-the-art summarization model by 9.68%, and even outperforms human-written references.[1]

1 Introduction

Every good article needs a good title, which should not only be able to condense the core meaning of the text, but also sound appealing to the readers for more exposure and memorableness. However, currently even the best Headline Generation (HG) system can only fulfill the above requirement yet performs poorly on the latter. For example, in Figure 1, the plain headline by an HG model "*Summ: Leopard Frog Found in New York City*" is less eye-catching than the style-carrying ones such as "*What's That Chuckle You Hear? It May Be the New Frog From NYC.*"



Original Headline:
Ribbit! Frog Species Found in New York City Has a Croak of Its Own

Article
New frog species discovered in New York City area. It has a distinctive croak, scientists find. Leopard frog species doesn't yet have a name.

HG Model Output: Summ: Leopard Frog Found in New York City

Humorous: What's that Chuckle You Hear? It May be the New Frog from NYC

Romantic: A New Frog with a Croak of Its Own Awaits its Name in the Roads of NYC

Click-Baity: 3 Facts about the New Frog with a Croak of Its Own

Figure 1: Given a news article, current HG models can only generate plain, factual headlines, failing to learn from the original human reference. It is also much less attractive than the headlines with humorous, romantic and click-baity styles.

To bridge the gap between the practical needs for attractive headlines and the plain HG by the current summarization systems, we propose a new task of Stylistic Headline Generation (SHG). Given an article, it aims to generate a headline with a target style such as humorous, romantic, and click-baity. It has broad applications in reader-adapted title generation, slogan suggestion, auto-fill for online post headlines, and many others.

SHG is a highly skilled creative process, and usually only possessed by expert writers. One of the most famous headlines in American publications, "*Sticks Nix Hick Pix,*" could be such an example. In contrast, the current best summarization systems are at most comparable to novice writers who provide a plain descriptive representation of the text body as the title (Cao et al., 2018b,a; Lin et al., 2018; Song et al., 2019; Dong et al., 2019). These systems usually use a language generation model that mixes styles with other linguistic patterns and inherently lacks a mechanism to control the style

[*] Corresponding author.

[1] Our code is available at https://github.com/jind11/TitleStylist.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5082–5093
July 5 - 10, 2020.

explicitly. More fundamentally, the training data comprise of a mixture of styles (e.g., the Gigaword dataset (Rush et al., 2017)), obstructing the models from learning a distinct style.

In this paper, we propose the new task SHG, to emphasize the explicit control of style in headline generation. We present a novel headline generation model, TitleStylist, to produce enticing titles with target styles including humorous, romantic, and click-baity. Our model leverages a multitasking framework to train both a summarization model on headline-article pairs, and a Denoising Autoencoder (DAE) on a style corpus. In particular, based on the transformer architecture (Vaswani et al., 2017), we use the style-dependent layer normalization and the style-guided encoder-attention to disentangle the language style factors from the text. This design enables us to use the shared content to generate headlines that are more relevant to the articles, as well as to control the style by plugging in a set of style-specific parameters. We validate the model on three tasks: humorous, romantic, and click-baity headline generation. Both automatic and human evaluations show that TitleStylist can generate headlines with the desired styles that appeal more to human readers, as in Figure 1.

The main contributions of our paper are listed below:

- To the best of our knowledge, it is the first research on the generation of attractive news headlines with styles without any supervised style-specific article-headline paired data.

- Through both automatic and human evaluation, we demonstrated that our proposed TitleStylist can generate relevant, fluent headlines with three styles (humor, romance, and clickbait), and they are even more attractive than human-written ones.

- Our model can flexibly incorporate multiple styles, thus efficiently and automatically providing humans with various creative headline options for references and inspiring them to think out of the box.

2 Related Work

Our work is related to summarization and text style transfer.

Headline Generation as Summarization

Headline generation is a very popular area of research. Traditional headline generation methods mostly focus on the extractive strategies using linguistic features and handcrafted rules (Luhn, 1958; Edmundson, 1964; Mathis et al., 1973; Salton et al., 1997; Jing and McKeown, 1999; Radev and McKeown, 1998; Dorr et al., 2003). To enrich the diversity of the extractive summarization, abstractive models were then proposed. With the help of neural networks, Rush et al. (2015) proposed attention-based summarization (ABS) to make Banko et al. (2000)'s framework of summarization more powerful. Many recent works extended ABS by utilizing additional features (Chopra et al., 2016; Takase et al., 2016; Nallapati et al., 2016; Shen et al., 2016, 2017a; Tan et al., 2017; Guo et al., 2017). Other variants of the standard headline generation setting include headlines for community question answering (Higurashi et al., 2018), multiple headline generation (Iwama and Kano, 2019), user-specific generation using user embeddings in recommendation systems (Liu et al., 2018), bilingual headline generation (Shen et al., 2018) and question-style headline generation (Zhang et al., 2018a).

Only a few works have recently started to focus on increasing the attractiveness of generated headlines (Fan et al., 2018; Xu et al., 2019). Fan et al. (2018) focuses on controlling several features of the summary text such as text length, and the style of two different news outlets, CNN and DailyMail. These controls serve as a way to boost the model performance, and the CNN- and DailyMail-style control shows a negligible improvement. Xu et al. (2019) utilized reinforcement learning to encourage the headline generation system to generate more sensational headlines via using the readers' comment rate as the reward, which however cannot explicitly control or manipulate the styles of headlines. Shu et al. (2018) proposed a style transfer approach to transfer a non-clickbait headline into a clickbait one. This method requires paired news articles-headlines data for the target style; however, for many styles such as humor and romance, there are no available headlines. Our model does not have this limitation, thus enabling transferring to many more styles.

Text Style Transfer

Our work is also related to text style transfer, which aims to change the style attribute of the text while preserving its content. First proposed by Shen et al. (2017b), it has achieved great progress in recent years (Xu et al., 2018; Lample et al., 2019; Zhang et al., 2018b; Fu et al., 2018; Jin et al., 2019; Yang et al., 2018; Jin et al., 2020). However, all these methods demand a text corpus for the target style; however, in our case, it is expensive and technically challenging to collect news headlines with humor and romance styles, which makes this category of methods not applicable to our problem.

3 Methods

3.1 Problem Formulation

The model is trained on a source dataset S and target dataset T. The source dataset $S = \{(\boldsymbol{a}^{(i)}, \boldsymbol{h}^{(i)})\}_{i=1}^{N}$ consists of pairs of a news article $\boldsymbol{a}$ and its *plain* headline $\boldsymbol{h}$. We assume that the source corpus has a distribution $P(A, H)$, where $A = \{\boldsymbol{a}^{(i)}\}_{i=1}^{N}$, and $H = \{\boldsymbol{h}^{(i)}\}_{i=1}^{N}$. The target corpus $T = \{\boldsymbol{t}^{(i)}\}_{i=1}^{M}$ comprises of sentences $\boldsymbol{t}$ written in a specific style (e.g., humor). We assume that it conforms to the distribution $P(T)$.

Note that the target corpus T only contains style-carrying sentences, not necessarily headlines — it can be just book text. Also no sentence $\boldsymbol{t}$ is paired with a news article. Overall, our task is to learn the conditional distribution $P(T|A)$ using only S and T. This task is fully unsupervised because there is *no* sample from the joint distribution $P(A, T)$.

3.2 Seq2Seq Model Architecture

For summarization, we adopt a sequence-to-sequence (Seq2Seq) model based on the Transformer architecture (Vaswani et al., 2017). As in Figure 2, it consists of a 6-layer encoder $E(\cdot; \boldsymbol{\theta_E})$ and a 6-layer decoder $G(\cdot; \boldsymbol{\theta_G})$ with a hidden size of 1024 and a feed-forward filter size of 4096. For better generation quality, we initialize with the MASS model (Song et al., 2019). MASS is pretrained by masking a sentence fragment in the encoder, and then predicting it in the decoder on large-scale English monolingual data. This pretraining is adopted in the current state-of-the-art systems across various summarization benchmark tasks including HG.

3.3 Multitask Training Scheme

To disentangle the latent style from the text, we adopt a multitask learning framework (Luong et al., 2015), training on summarization and DAE simultaneously (as shown in Figure 3).

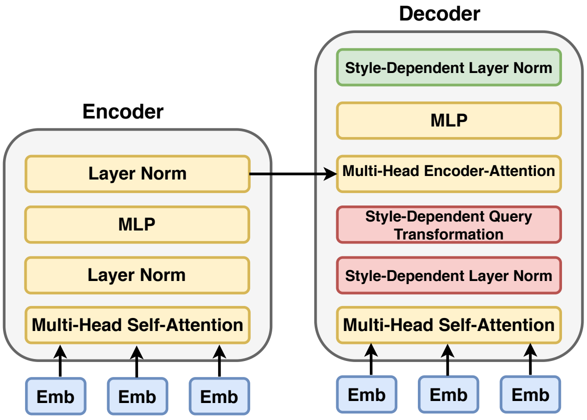

Decoder
Style-Dependent Layer Norm
MLP
Multi-Head Encoder-Attention
Style-Dependent Query Transformation
Style-Dependent Layer Norm
Multi-Head Self-Attention
Emb
Emb
Emb
Encoder
Layer Norm
MLP
Layer Norm
Multi-Head Self-Attention
Emb
Emb
Emb

Figure 2: The Transformer-based architecture of our model.

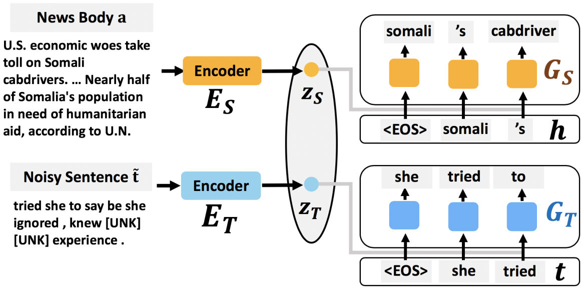

News Body a
U.S. economic woes take toll on Somali cabdrivers. ... Nearly half of Somalia's population in need of humanitarian aid, according to U.N.
Encoder
E_S
Noisy Sentence $\tilde{t}$
tried she to say be she ignored , knew [UNK] [UNK] experience .
Encoder
E_T
z_S
z_T
somali
's
cabdriver
G_S
<EOS>
somali
's
h
she
tried
to
G_T
<EOS>
she
tried
t

Figure 3: Training scheme. Multitask training is adopted to combine the summarization and DAE tasks.

Supervised Seq2Seq Training for E_S and G_S With the source domain dataset S, based on the encoder-decoder architecture, we can learn the conditional distribution $P(H|A)$ by training $\boldsymbol{z_S} = E_S(A)$ and $H_S = G_S(\boldsymbol{z_S})$ to solve the supervised Seq2Seq learning task, where $\boldsymbol{z_S}$ is the learned latent representation in the source domain. The loss function of this task is

$$\mathcal{L}_S(\boldsymbol{\theta_{E_S}}, \boldsymbol{\theta_{G_S}}) = \mathbb{E}_{(\boldsymbol{a},\boldsymbol{h})\sim S}[-\log p(\boldsymbol{h}|\boldsymbol{a}; \boldsymbol{\theta_{E_S}}, \boldsymbol{\theta_{G_S}})], \tag{1}$$

where $\boldsymbol{\theta_{E_S}}$ and $\boldsymbol{\theta_{G_S}}$ are the set of model parameters of the encoder and decoder in the source domain and $p(\boldsymbol{h}|\boldsymbol{a})$ denotes the overall probability of generating an output sequence $\boldsymbol{h}$ given the input article $\boldsymbol{a}$, which can be further expanded as follows:

$$p(\boldsymbol{h}|\boldsymbol{a}; \boldsymbol{\theta_{E_S}}, \boldsymbol{\theta_{G_S}}) = \prod_{t=1}^{L} p(h_t|\{h_1, ..., h_{t-1}\}, \boldsymbol{z_S}; \boldsymbol{\theta_{G_S}}), \tag{2}$$

where L is the sequence length.

DAE Training for $\boldsymbol{\theta_{E_T}}$ and $\boldsymbol{\theta_{G_T}}$ For the target style corpus T, since we only have the sentence $\boldsymbol{t}$ without paired news articles, we train $\boldsymbol{z_T} = E_T(\tilde{\boldsymbol{t}})$ and $\boldsymbol{t} = G_T(\boldsymbol{z_T})$ by solving an unsupervised reconstruction learning task, where $\boldsymbol{z_T}$ is the learned latent representation in the target domain, and $\tilde{\boldsymbol{t}}$ is the corrupted version of $\boldsymbol{t}$ by randomly deleting or blanking some words and shuffling the word orders. To train the model, we minimize the reconstruction error $\mathcal{L}_T$:

$$\mathcal{L}_T(\boldsymbol{\theta_{E_T}}, \boldsymbol{\theta_{G_T}}) = \mathbb{E}_{\boldsymbol{t}\sim\boldsymbol{T}}[-\log p(\boldsymbol{t}|\tilde{\boldsymbol{t}})], \tag{3}$$

where $\boldsymbol{\theta_{E_T}}$ and $\boldsymbol{\theta_{G_T}}$ are the set of model parameters for the encoder and generator in the target domain. We train the whole model by jointly minimizing the supervised Seq2Seq training loss $\mathcal{L}_S$ and the unsupervised denoised auto-encoding loss $\mathcal{L}_T$ via multitask learning, so the total loss becomes

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta_{E_S}}, \boldsymbol{\theta_{G_S}}, \boldsymbol{\theta_{E_T}}, \boldsymbol{\theta_{G_T}}) &= \lambda\mathcal{L}_S(\boldsymbol{\theta_{E_S}}, \boldsymbol{\theta_{G_S}}) \\ &+ (1-\lambda)\mathcal{L}_T(\boldsymbol{\theta_{E_T}}, \boldsymbol{\theta_{G_T}}),\end{aligned} \tag{4}$$

where λ is a hyper-parameter.

3.4 Parameter-Sharing Scheme

More constraints are necessary in the multitask training process. We aim to infer the conditional distribution as $P(T|A) = G_T(E_S(A))$. However, without samples from $P(A,T)$, this is a challenging or even impossible task if E_S and E_T, or G_S and G_T are completely independent of each other. Hence, we need to add some constraints to the network by relating E_S and E_T, and G_S and G_T. The simplest design is to share all parameters between E_S and E_T, and apply the same strategy to G_S and G_T. The intuition behind this design is that by exposing the model to both summarization task and style-carrying text reconstruction task, the model would acquire some sense of the target style while summarizing the article. However, to encourage the model to better disentangle the content and style of text and more explicitly learn the style contained in the target corpus T, we share all parameters of the encoder between two domains, i.e., between E_S and E_T, whereas we divide the parameters of the decoder into two types: style-independent parameters $\boldsymbol{\theta_{\text{ind}}}$ and style-dependent parameters $\boldsymbol{\theta_{\text{dep}}}$. This means that only the style-independent parameters are shared between G_S and G_T while the style-dependent parameters are not. More specifically, the parameters of the layer normalization and encoder attention modules are made style-dependent as detailed below.

Type 1. Style Layer Normalization Inspired by previous work on image style transfer (Dumoulin et al., 2016), we make the scaling and shifting parameters for layer normalization in the transformer architecture un-shared for each style. This *style layer normalization* approach aims to transform a layer's activation $\boldsymbol{x}$ into a normalized activation $\boldsymbol{z}$ specific to the style s:

$$\boldsymbol{z} = \gamma_s(\frac{\boldsymbol{x}-\mu}{\sigma}) - \beta_s, \tag{5}$$

where μ and σ are the mean and standard deviation of the batch of $\boldsymbol{x}$, and γ_s and β_s are style-specific parameters learned from data.

Specifically, for the transformer decoder architecture, we use a style-specific self-attention layer normalization and final layer normalization for the source and target domains on all six decoder layers.

Type 2. Style-Guided Encoder Attention Our model architecture contains the attention mechanism, where the decoder infers the probability of the next word not only conditioned on the previous words but also on the encoded input hidden states. The attention patterns should be different for the summarization and the reconstruction tasks due to their different inherent nature. We insert this thinking into the model by introducing the *style-guided encoder attention* into the multi-head attention module, which is defined as follows:

$$\boldsymbol{Q} = \textbf{query} \cdot \boldsymbol{W_q^s} \tag{6}$$

$$\boldsymbol{K} = \textbf{key} \cdot \boldsymbol{W_k} \tag{7}$$

$$\boldsymbol{V} = \textbf{value} \cdot \boldsymbol{W_v} \tag{8}$$

$$\text{Att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{QK}^{\textbf{tr}}}{\sqrt{d_{\text{model}}}}\right)\boldsymbol{V}, \tag{9}$$

where **query**, **key**, and **value** denote the triple of inputs into the multi-head attention module; $\boldsymbol{W_q^s}$, $\boldsymbol{W_k}$, and $\boldsymbol{W_v}$ denote the scaled dot-product matrix for affine transformation; d_{model} is the dimension of the hidden states. We specialize the dot-product matrix $\boldsymbol{W_q^s}$ of the query for different styles, so that $\boldsymbol{Q}$ can be different to induce diverse attention patterns.

4 Experiments

4.1 Datasets

We compile a rich source dataset by combining the New York Times (NYT) and CNN, as well as three target style corpora on humorous, romantic, and click-baity text. The average sentence length in the NYT, CNN, Humor, Romance, and Clickbait datasets are 8.8, 9.2, 12.6, 11.6 and 8.7 words, respectively.

4.1.1 Source Dataset

The source dataset contains news articles paired with corresponding headlines. To enrich the training corpus, we combine two datasets: the New York Times (56K) and CNN (90K). After combining these two datasets, we randomly selected 3,000 pairs as the validation set and another 3,000 pairs as the test set.

We first extracted the archival abstracts and headlines from the New York Times (NYT) corpus (Sandhaus, 2008) and treat the abstracts as the news articles. Following the standard pre-processing procedures (Kedzie et al., 2018),[2] we filtered out advertisement-related articles (as they are very different from news reports), resulting in 56,899 news abstracts-headlines pairs.

We then add into our source set the CNN summarization dataset, which is widely used for training abstractive summarization models (Hermann et al., 2015).[3] We use the short summaries in the original dataset as the news abstracts and automatically parsed the headlines for each news from the dumped news web pages,[4] and in total collected 90,236 news abstract-headline pairs.

4.1.2 Three Target Style Corpora

Humor and Romance For the target style datasets, we follow (Chen et al., 2019) to use humor and romance novel collections in BookCorpus (Zhu et al., 2015) as the Humor and Romance datasets.[5] We split the documents into sentences, tokenized the text, and collected 500K sentences as our datasets.

Clickbait We also tried to learn the writing style from the click-baity headlines since they have shown superior attraction to readers. Thus we used *The Examiner - SpamClickBait News* dataset, denoted as the Clickbait dataset.[6] We collected 500K headlines for our use.

Some examples from each style corpus are listed in Table 1.

Style	Examples
Humor	- The crowded beach like houses in the burbs and the line ups at Walmart. - Berthold stormed out of the brewing argument with his violin and bow and went for a walk with it to practice for the much more receptive polluted air.
Romance	- "I can face it joyously and with all my heart, and soul!" she said. - With bright blue and green buttercream scales, sparkling eyes, and purple candy melt wings, it sat majestically on a rocky ledge made from chocolate.
Clickbait	- 11-Year-Old Girl and 15-Year-Old Boy Accused of Attempting to Kill Mother: Who Is the Adult? - Chilly, Dry Weather Welcomes 2010 to South Florida - End Segregation in Alabama-Bryce Hospital Sale Offers a Golden Opportunity

Table 1: Examples of three target style corpora: humor, romance, and clickbait.

4.2 Baselines

We compared the proposed TitleStylist against the following five strong baseline approaches.

Neural Headline Generation (NHG) We train the state-of-the-art summarization model, MASS (Song et al., 2019), on our collected news abstracts-headlines paired data.

Gigaword-MASS We test an off-the-shelf headline generation model, MASS from (Song et al., 2019), which is already trained on Gigaword, a large-scale headline generation dataset with around 4 million articles.[7]

Neural Story Teller (NST) It breaks down the task into two steps, which first generates headlines from the aforementioned NHG model, then applies style shift techniques to generate style-specific headlines (Kiros et al., 2015). In brief, this method uses the Skip-Thought model to encode a sentence into a representation vector and then manipulates its style by a linear transformation. Afterward, this transformed representation vector is used to initialize a language model pretrained on a style-specific corpus so that a stylistic headline can be generated. More details of this method can refer to the official website.[8]

[2] https://github.com/kedz/summarization-datasets

[3] We use CNN instead of the DailyMail dataset since DailyMail headlines are very long and more like short summaries.

[4] https://cs.nyu.edu/~kcho/DMQA/

[5] https://www.smashwords.com/

[6] https://www.kaggle.com/therohk/examine-the-examiner

[7] https://github.com/harvardnlp/sent-summary

[8] https://github.com/ryankiros/neural-storyteller

Fine-Tuned We first train the NHG model as mentioned above, then further fine-tuned it on the target style corpus via DAE training.

Multitask We share all parameters between E_S and E_T, and between G_S and G_T, and trained the model on both the summarization and DAE tasks. The model architecture is the same as NHG.

4.3 Evaluation Metrics

To evaluate the performance of the proposed TitleStylist in generating attractive headlines with styles, we propose a comprehensive twofold strategy of both automatic evaluation and human evaluation.

4.3.1 Setup of Human Evaluation

We randomly sampled 50 news abstracts from the test set and asked three native-speaker annotators for evaluation to score the generated headlines. Specifically, we conduct two tasks to evaluate on four criteria: (1) relevance, (2) attractiveness, (3) language fluency, and (4) style strength. For the first task, the human raters are asked to evaluate these outputs on the first three aspects, relevance, attractiveness, and language fluency on a Likert scale from 1 to 10 (integer values). For **relevance**, human annotators are asked to evaluate how semantically relevant the headline is to the news body. For **attractiveness**, annotators are asked how attractive the headlines are. For **fluency**, we ask the annotators to evaluate how fluent and readable the text is. After the collection of human evaluation results, we averaged the scores as the final score. In addition, we have another independent human evaluation task about the **style strength** – we present the generated headlines from TitleStylist and baselines to the human judges and let them choose the one that most conforms to the target style such as humor. Then we define the style strength score as the proportion of choices.

4.3.2 Setup of Automatic Evaluation

Apart from the comprehensive human evaluation, we use automatic evaluation to measure the generation quality through two conventional aspects: summarization quality and language fluency. Note that the purpose of this two-way automatic evaluation is to confirm that the performance of our model is in an acceptable range. Good automatic evaluation performances are necessary proofs to compliment human evaluations on the model effectiveness.

Summarization Quality We use the standard automatic evaluation metrics for summarization with the original headlines as the reference: BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015). For ROUGE, we used the Files2ROUGE[9] toolkit, and for other metrics, we used the pycocoeval toolkit.[10]

Language Fluency We fine-tuned the GPT-2 medium model (Radford et al., 2019) on our collected headlines and then used it to measure the perplexity (PPL) on the generated outputs.[11]

4.4 Experimental Details

We used the fairseq code base (Ott et al., 2019). During training, we use Adam optimizer with an initial learning rate of 5×10^{-4}, and the batch size is set as 3072 tokens for each GPU with the parameters update frequency set as 4. For the random corruption for DAE training, we follow the standard practice to randomly delete or blank the word with a uniform probability of 0.2, and randomly shuffled the word order within 5 tokens. All datasets are lower-cased. λ is set as 0.5 in experiments. For each iteration of training, we randomly draw a batch of data either from the source dataset or from the target style corpus, and the sampling strategy follows the uniform distribution with the probability being equal to λ.

5 Results and Discussion

5.1 Human Evaluation Results

The human evaluation is to have a comprehensive measurement of the performances. We conduct experiments on four criteria, relevance, attraction, fluency, and style strength. We summarize the human evaluation results on the first three criteria in Table 2, and the last criteria in Table 4. Note that through automatic evaluation, the baselines NST, Fine-tuned, and Gigaword-MASS perform poorer than other methods (in Section 5.2), thereby we removed them in human evaluation to save unnecessary work for human raters.

Relevance We first look at the relevance scores in Table 2. It is interesting but not surprising that the pure summarization model NHG achieves the highest relevance score. The outputs from NHG

[9] `https://github.com/pltrdy/files2rouge`
[10] `https://github.com/Maluuba/nlg-eval`
[11] PPL on the development set is 42.5

Style	Settings	Relevance	Attraction	Fluency
None	NHG	**6.21**	8.47	9.31
	Human	5.89	8.93	9.33
Humor	Multitask	5.51	8.61	9.11
	TitleStylist	5.87	8.93	9.29
Romance	Multitask	5.67	8.54	8.91
	TitleStylist	5.86	8.87	9.14
Clickbait	Multitask	5.67	8.71	9.21
	TitleStylist	5.83	**9.29**	**9.44**

Table 2: Human evaluation on three aspects: relevance, attraction, and fluency. “None” represents the original headlines in the dataset.

are usually like an organic reorganization of several keywords in the source context (as shown in Table 3), thus appearing most relevant. It is noteworthy that the generated headlines of our TitleStylist for all three styles are close to the original human-written headlines in terms of relevance, validating that our generation results are qualified in this aspect. Another finding is that more attractive or more stylistic headlines would lose some relevance since they need to use more words outside the news body for improved creativity.

Attraction In terms of attraction scores in Table 2, we have three findings: (1) The human-written headlines are more attractive than those from NHG, which agrees with our observation in Section 1. (2) Our TitleStylist can generate more attractive headlines over the NHG and Multitask baselines for all three styles, demonstrating that adapting the model to these styles could improve the attraction and specialization of some parameters in the model for different styles can further enhance the attraction. (3) Adapting the model to the “Clickbait” style could create the most attractive headlines, even out-weighting the original ones, which agrees with the fact that click-baity headlines are better at drawing readers’ attention. To be noted, although we learned the “Clickbait” style into our summarization system, we still made sure that we are generating relevant headlines instead of too exaggerated ones, which can be verified by our relevance scores.

Fluency The human-annotated fluency scores in Table 2 verified that our TitleStylist generated headlines are comparable or superior to the human-written headlines in terms of readability.

Style Strength We also validated that our TitleStylist can carry more styles compared with the Multitask and NHG baselines by summarizing the percentage of choices by humans for the most humorous or romantic headlines in Table 4.

5.2 Automatic Evaluation Results

Apart from the human evaluation of the overall generation quality on four criteria, we also conducted a conventional automatic assessment to gauge only the summarization quality. This evaluation does not take other measures such as the style strength into consideration, but it serves as important complimentary proof to ensure that the model has an acceptable level of summarization ability.

Table 5 summarizes the automatic evaluation results of our proposed TitleStylist model and all baselines. We use the summarization-related evaluation metrics, i.e., BLEU, ROUGE, CIDEr, and METEOR, to measure how relevant the generated headlines are to the news articles, to some extent, by comparing them to the original human-written headlines. In Table 5, the first row “NHG” shows the performance of the current state-of-the-art summarization model on our data, and Table 3 provides two examples of its generation output. Our ultimate goal is to generate more attractive headlines than these while maintaining relevance to the news body.

From Table 5, the baseline Gigaword-MASS scored worse than NHG, revealing that directly applying an off-the-shelf headline generation model to new in-domain data is not feasible, although this model has been trained on more than 20 times larger dataset. Both NST and Fine-tuned baselines present very poor summarization performance, and the reason could be that both of them cast the problem into two steps: summarization and style transfer, and the latter step is absent of the summarization task, which prevents the model from maintaining its summarization capability.

In contrast, the Multitask baseline involves the summarization and style transfer (via reconstruction training) processes at the same time and shows superior summarization performance even compared with NHG. This reveals that the unsupervised reconstruction task can indeed help improve the supervised summarization task. More importantly, we use two different types of corpora for the reconstruction task: one consists of headlines that are similar to the news data for the summarization task, and the other consists of text from novels that are entirely different from the news data. However,

News Abstract	Turkey's bitter history with Kurds is figuring prominently in its calculations over how to deal with Bush administration's request to use Turkey as the base for thousands of combat troops if there is a war with Iraq; Recep Tayyip Erdogan, leader of Turkey's governing party, says publicly for the first time that future of Iraq's Kurdish area, which abuts border region of Turkey also heavily populated by Kurds, is weighing heavily on negotiations; Hints at what Turkish officials have been saying privately for weeks: if war comes to Iraq, overriding Turkish objective would be less helping Americans topple Saddam Hussein, but rather preventing Kurds in Iraq from forming their own state.	Reunified Berlin is commemorating 40th anniversary of the start of construction of Berlin wall, almost 12 years since Germans jubilantly celebrated reopening between east and west and attacked hated structure with sledgehammers; Some Germans are championing the preservation of wall at the time when little remains beyond few crumbling remnants to remind Berliners of unhappy division that many have since worked hard to heal and put behind them; What little remains of physical wall embodies era that Germans have yet to resolve for themselves; They routinely talk of 'wall in the mind' to describe social and cultural differences that continue to divide easterners and westerners.
Human	Turkey assesses question of Kurds	The wall Berlin can't quite demolish
NHG	Turkey's bitter history with Kurds	Construction of Berlin wall is commemorated
Humor	What if there is a war with Kurds?	The Berlin wall, 12 years later, is still there?
Romance	What if the Kurds say "No" to Iraq?	The Berlin wall: from the past to the present
Clickbait	For Turkey, a long, hard road	East vs West, Berlin wall lives on

Table 3: Examples of style-carrying headlines generated by TitleStylist.

Style	**NHG**	**Multitask**	**TitleStylist**
Humor	18.7	35.3	**46.0**
Romance	24.7	34.7	**40.6**
Clickbait	13.8	35.8	**50.4**

Table 4: Percentage of choices (%) for the most humorous or romantic headlines among TitleStylist and two baselines NHG and Multitask.

unsupervised reconstruction training on both types of data can contribute to the summarization task, which throws light on the potential future work in summarization by incorporating unsupervised learning as augmentation.

We find that in Table 5 TitleStylist-F achieves the best summarization performance. This implicates that, compared with the Multitask baseline where the two tasks share all parameters, specialization of layer normalization and encoder-attention parameters can make G_S focus more on summarization.

It is noteworthy that the summarization scores for TitleStylist are lower than TitleStylist-F but still comparable to NHG. This agrees with the fact that the G_T branch more focuses on bringing in stylistic linguistic patterns into the generated summaries, thus the outputs would deviate from the pure summarization to some degree. However, the relevance degree of them remains close to the baseline NHG, which is the starting point we want to improve on. Later in the next section, we will further validate that these headlines are faithful to the new article through human evaluation.

We also reported the perplexity (PPL) of the generated headlines to evaluate the language fluency, as shown in Table 5. All outputs from baselines NHG and Multitask and our proposed TitleStylist show similar PPL compared with the test set (used in the fine-tuning stage) PPL 42.5, indicating that they are all fluent expressions for news headlines.

5.3 Extension to Multi-Style

We progressively expand TitleStylist to include all three target styles (humor, romance, and clickbait) to demonstrate the flexibility of our model. That is, we simultaneously trained the summarization task on the headlines data and the DAE task on the three target style corpora. And we made the layer normalization and encoder-attention parameters specialized for these four styles (fact, humor, romance, and clickbait) and shared the other parameters. We compared this multi-style version, TitleStylist-Versatile, with the previously presented single-style counterpart, as shown in Table 6. From this table, we see that the BLEU and ROUGE-L scores of TitleStylist-Versatile are comparable to TitleStylist for all three styles. Besides, we conducted another human study to determine the better headline between the two models in terms of attraction, and we allow human annotators to choose both options if they deem them as equivalent. The result is presented in the last column of Table 6, which shows that the attraction of TitleStylist-Versatile outputs is competitive to TitleStylist. TitleStylist-Versatile thus generates multiple headlines in different styles altogether, which is a novel and efficient

Style Corpus	Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	CIDEr	METEOR	PPL (↓)	Len. Ratio (%)
None	NHG	12.9	27.7	9.7	24.8	0.821	0.123	40.4	8.9
	Gigaword-MASS	9.2	22.6	6.4	20.1	0.576	0.102	65.0	9.7
Humor	NST	5.8	17.8	4.3	16.1	0.412	0.078	361.3	9.2
	Fine-tuned	4.3	15.7	3.4	13.2	0.140	0.093	398.8	3.9
	Multitask	14.7	28.9	**11.6**	26.1	0.995	0.134	40.0	9.5
	TitleStylist	13.3	28.1	10.3	25.4	0.918	0.127	46.2	10.6
	TitleStylist-F	**15.2**	**29.2**	**11.6**	**26.3**	**1.022**	**0.135**	**39.3**	9.7
Romance	NST	2.9	9.8	0.9	9.0	0.110	0.047	434.1	6.2
	Fine-tuned	5.1	18.7	4.5	16.1	0.023	0.128	132.2	2.8
	Multitask	14.8	28.7	11.5	25.9	0.997	0.132	40.5	9.7
	TitleStylist	12.0	27.2	10.1	24.4	0.832	0.134	40.1	7.4
	TitleStylist-F	**15.0**	**29.0**	**11.7**	**26.2**	**1.005**	**0.134**	**39.0**	9.8
Clickbait	NST	2.5	8.4	0.6	7.8	0.089	0.041	455.4	6.3
	Fine-tuned	4.7	17.3	4.0	15.0	0.019	0.116	172.0	2.8
	Multitask	14.5	28.3	11.2	25.5	0.980	0.132	**38.5**	9.7
	TitleStylist	11.5	26.6	9.8	23.7	0.799	**0.134**	40.7	7.3
	TitleStylist-F	**14.7**	**28.6**	**11.4**	**25.9**	**0.981**	0.133	38.9	9.6

Table 5: Automatic evaluation results of our TitleStylist and baselines. The test set of each style is the same, but the training set is different depending on the target style as shown in the "Style Corpus" column. "None" means no style-specific dataset, and "Humor", "Romance" and "Clickbait" corresponds to the datasets we introduced in Section 4.1.2. During the inference phase, our TitleStylist can generate two outputs: one from G_T and the other from G_S. Outputs from G_T are style-carrying, so we denote it as "TitleStylist"; outputs from G_S are plain and factual, thus denoted as "TitleStylist-F." The last column "Len. Ratio" denotes the average ratio of abstract length to the generated headline length by the number of words.

Style	Model	BLEU	RG-L	Pref. (%)
None	TitleStylist-Versatile	14.5	25.8	—
Humor	TitleStylist-Versatile	12.3	24.5	42.6
	TitleStylist	**13.3**	**25.4**	**57.4**
Romance	TitleStylist-Versatile	**12.0**	24.2	46.3
	TitleStylist	**12.0**	**24.4**	**53.7**
Clickbait	TitleStylist-Versatile	**13.1**	**24.9**	**52.9**
	TitleStylist	11.5	23.7	47.1

Table 6: Comparison between TitleStylist-Versatile and TitleStylist. "RG-L" denotes ROUGE-L, and "Pref." denotes preference.

feature.

6 Conclusion

We have proposed a new task of Stylistic Headline Generation (SHG) to emphasize explicit control of styles in headline generation for improved attraction. To this end, we presented a multitask framework to induce styles into summarization, and proposed the parameters sharing scheme to enhance both summarization and stylization capabilities. Through experiments, we validated our proposed TitleStylist can generate more attractive headlines than state-of-the-art HG models.

Acknowledgement

We appreciate all the volunteer native speakers (Shreya Karpoor, Lisa Orii, Abhishek Mohan, Paloma Quiroga, etc.) for the human evaluation of our study, and thank the reviewers for their inspiring comments. Joey Tianyi Zhou is partially supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A1b0045).

References

Michele Banko, Vibhu O Mittal, and Michael J Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 318–325. Association for Computational Linguistics.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018a. Retrieve, rerank and rewrite: Soft template based neural summarization. In *ACL*.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018b. Faithful to the original: Fact aware neural abstractive summarization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Cheng-Kuan Chen, Zhu Feng Pan, Ming-Yu Liu, and Min Sun. 2019. Unsupervised stylish image description generation via domain layer norm. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8151–8158. AAAI Press.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

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*, pages 1–8. Association for Computational Linguistics.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2016. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*.

HP Edmundson. 1964. Problems in automatic abstracting. *Communications of the ACM*, 7(4):259–263.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 45–54. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yidi Guo, Heyan Huang, Yang Gao, and Chi Lu. 2017. Conceptual multi-layer neural network model for headline generation. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 355–367. Springer.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Tatsuru Higurashi, Hayato Kobayashi, Takeshi Masuyama, and Kazuma Murao. 2018. Extractive headline generation based on learning to rank for community question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1742–1753.

Kango Iwama and Yoshinobu Kano. 2019. Multiple news headlines generation using page metadata. In *Proceedings of the 12th International Conference on Natural Language Generation, 2019*. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Unsupervised domain adaptation for neural machine translation with iterative back translation. *arXiv preprint arXiv:2001.08140*.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. Unsupervised text attribute transfer via iterative matching and translation. In *IJCNLP 2019*.

Hongyan Jing and Kathleen McKeown. 1999. The decomposition of human-written summary sentences.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. *arXiv preprint arXiv:1810.12343*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *ICLR*.

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

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *ACL*.

Tianshang Liu, Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Review headline generation with user embedding. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data - 17th China National Conference, CCL 2018, and 6th International Symposium, NLP-NABD 2018, Changsha, China, October 19-21, 2018, Proceedings*, pages 324–334.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multitask sequence to sequence learning. *CoRR*, abs/1511.06114.

Betty A Mathis, James E Rush, and Carol E Young. 1973. Improvement of automatic abstracts by the use of structural analysis. *Journal of the American Society for Information Science*, 24(2):101–109.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Dragomir R Radev and Kathleen R McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):470–500.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Alexander M Rush, SEAS Harvard, Sumit Chopra, and Jason Weston. 2017. A neural attention model for sentence summarization. In *ACLWeb. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic text structuring and summarization. *Information processing & management*, 33(2):193–207.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

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.

Shiqi Shen, Yun Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 26(12):2319–2327.

Shiqi Shen, Yu Zhao, Zhiyuan Liu, Maosong Sun, et al. 2016. Neural headline generation with sentence-wise optimization. *arXiv preprint arXiv:1604.01904*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017b. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6830–6841.

Kai Shu, Suhang Wang, Thai Le, Dongwon Lee, and Huan Liu. 2018. Deep headline generation for clickbait detection. *2018 IEEE International Conference on Data Mining (ICDM)*, pages 467–476.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

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.

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.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *ACL*.

Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? sensational headline generation with auto-tuned reinforcement learning. *ArXiv*, abs/1909.03582.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P. Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 7298–7309.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Huanhuan Cao, and Xueqi Cheng. 2018a. Question headline generation for news articles. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 617–626.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018b. Style transfer as unsupervised machine translation. *ArXiv*, abs/1808.07894.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.