Generating summaries tailored to target characteristics

Kushal Chawla[‡], Hrituraj Singh[†], Arijit Pramanik[*], Mithlesh Kumar[#],
Balaji Vasan Srinivasan[‡]

[‡]Adobe Research, Bangalore, India
[†]Indian Institute of Technology, Rourkee
[*]Indian Institute of Technology, Bombay
[#]Indian Institute of Technology, Kanpur
kchawla@adobe.com,balsrini@adobe.com

Abstract. Recently, research efforts have gained pace to cater to varied user preferences while generating text summaries. While there have been attempts to incorporate a few handpicked characteristics such as length or entities, a holistic view around these preferences is missing and crucial insights on why certain characteristics should be incorporated in a specific manner are absent. With this objective, we provide a categorization around these characteristics relevant to the task of text summarization: one, focusing on what content needs to be generated and second, focusing on the stylistic aspects of the output summaries. We use our insights to provide guidelines on appropriate methods to incorporate various classes characteristics in sequence-to-sequence summarization framework. Our experiments with incorporating topics, readability and simplicity indicate the viability of the proposed prescriptions.

1 Introduction

Automatic text summarization [1] is the task of generating a summary of an input document while retaining the key aspects. Such a summary helps in presenting the important content from a long input text in a succinct form for quick information consumption. Traditional methods for summarization [1] extract key sentences from the source text to construct an 'extractive' summary. Recent efforts towards 'abstractive' summarization have geared towards generating human-like, paraphrased summaries from the input article [2, 3].

While these algorithms allow for the generation of a single summary, it is often desirable to generate summary variants tailored towards specific characteristics. For instance, readers may be interested in summaries of different lengths or might want to focus on specific entities/topics from the input text [4–6]. Depending on different age groups of the readers, they might prefer formal/informal variants of the summary [7]. Irrespective of the application context, it has been shown that incorporating these characteristics at the time of generation can yield more contextual summaries [8].

Recent works have proposed different ways to incorporate target characteristics at the time of summary generation: introducing modifications to the

dataset [4, 5], architecture [7], learning objectives [9] or the decoder probabilities [8]. However, all these attempts handpick a few characteristics and propose ways to incorporate them. In the absence of appropriate insights, it is unclear as to why these methodologies work in tuning the summaries towards the chosen characteristics and why the same cannot be extended for other characteristics. In this work, our objective is to gain a holistic understanding around these additional constraints, centering on the task of text summarization.

Taking a step in this direction, we propose a categorization of these characteristics into 1) **content-specific**, which primarily focus on what content is presented pivoting on the semantics or information presented in the output summary and 2) **style-specific** focusing on the stylistic expressions pivoting on the linguistic presentation in the output summary. Through a comparative evaluation of various existing and proposed methods, we further prescribe guidelines to help choosing the right framework for tailoring these categories of characteristics. Our primary contribution is providing a categorization of target characteristics as content and style specific, towards a holistic understanding of tailored summary generation. Additionally, we propose an attention-boosting approach to improve tailoring of content-specific characteristics and a policy gradient based algorithm to incorporate stylistic characteristics in summaries.

2 Related Work

Traditional methods for summarization [1] extract key sentences from the source text to construct an 'extractive' summary. Features like descriptiveness of words and word frequencies have been explored to choose the summary sentences. However, humans summarize by understanding the content and paraphrasing the understood content into a summary. Extractive summarization is hence unable to produce 'human-like' summaries. This has led to efforts towards 'abstractive' summarization, which paraphrase summaries from input article content.

Early attempts at abstractive summarization created summary sentences either based on templates [10, 11] or used ILP-based sentence compression techniques [12–14]. With the advent of deep sequence to sequence models [15], attention-based neural models have been proposed for summarizing long sentences [16, 17]. These approaches were further improved by incorporating Abstract Meaning Representations [18] and using hierarchical encoding networks [19]. More recent approaches [2, 3] have focused on large scale datasets for summarization such as CNN/DailyMail corpus [20, 19]. Gulcehre et al. [21] introduced the ability to copy out-of-vocabulary words from the article to incorporate rarely seen words like names in the generated text. Tu et al. [22] included the concept of coverage, to prevent the models from repeating the same phrases while generating a sentence. See et al. [3] proposed a pointer-generator framework which incorporates these improvements, and also learns to switch between generating new words and copying words from the source article.

Although research has primarily focused on unconditioned abstractive text summarization, there have been some recent efforts to **incorporate a variety of additional constraints** into the generation algorithm. Fan et al. [4] use explicit

indicators to control the length of the output summary, imposing a constraint on how detailed the output needs to be. They extend the same technique to also control the entities which must be focused on while generating the summary. Krishna et al. [5] generated topic-oriented summaries by using an indicator topic-vector along with the input representation. Each of these approaches aim to control the *information* presented in the generated summary and therefore modifies the attention distribution to focus on appropriate parts of the text as dictated by the target characteristics. We group these approaches as **content-specific** characteristic tuning.

Another direction of efforts have attempted to incorporate aspects like sentiments or tense using generative models like variational auto-encoders [23] or using adversarial training [24]. Focusing on politeness, Sennrich et al. [25] propose modifications to neural machine translation setup to generate polite variants. Ficler et al. [26] propose the use of a conditional language model to generate text with variations such as descriptiveness, personal and sentiment simultaneously. More recently, generating text with varying levels of formality was studied in Machine Translation [7, 27]. Oraby et al. [28] attempt to control personality dimensions in generation, namely, agreeable, disagreeable, conscientious, unconscientious and extravert by using indicator tokens and stylistic encodings. Krishna et al. [8] modify the decoding algorithm to produce readable and simple summaries. Each of these exploration focus primarily of tuning the linguistic presentation of the generated text and we group these as **style-specific** characteristic tuning.

Policy learning based approaches [9] have shown promise to control several qualitative characteristics explicitly and can potentially be used for both content and style specific characteristics. While it has been successfully deployed for metrics like ROUGE [9], we show its applicability to style-specific characteristics by proposing a policy gradient framework for readability and simplicity.

Content vs Style: The notion of style and the associated nomenclature is quite convoluted in the literature [29]. Approaches in style transfer try to obtain independent latent representations for style and semantics [30],[31], [24], [32], [33], [34] of the content. Using this interpretation, we see the output of a text generation system as a combination of the semantics of the information which is being presented and the style associated with that content. However, unlike these approaches that learn the notion of style implicitly from the available corpora, we fragment style across a set of dimensions such as readability, simplicity, etc. These are referred to as various aspects of style aligning with ad-hoc approaches towards style transfer as described in [29].

3 Pointer-Generator Framework

We base all our explorations on the pointer generator network [3]. However, all our findings and insights are generic and can be extended to any other frameworks without loss of generality. We describe the pointer generator framework here for the sake of completion and please refer to [3] for more details on the framework. The pointer generator network [3] consists of an encoder and a de-

coder, both based on LSTM architecture. Given an input article, the encoder takes the embedding vectors of each word in the source text A and computes the encoder hidden states $h_1, h_2, ..h_n$. The final hidden state is passed to a decoder, which computes a hidden state s_t at each decoding step and calculates an attention distribution a^t over all words as $a^t = \text{softmax}(e^t)$, where,

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{att}) \tag{1}$$

where v, W_h, W_s and b_{att} are model parameters to be trained. The attention is a probability distribution over words in the source text, which aids the decoder in generating the next word in the summary using words in the source text with higher attention. The context vector $h_t^* = \sum_{i=1}^{n} a_i^t h_i$, is a weighted sum of the encoder hidden states (weighted on a_i^t the attention on the ith input word at t^{th} step) and is used to determine the next word to be generated. The attention distribution allows the network to focus on specific parts of the input as the output summary is generated. To tailor a summary to various content-specific characteristics, it is important to modify this attention distribution to focus on the appropriate parts of the input text as required by the characteristics tuned.

At each decoding step, the decoder also gets the last word y_t in the summary generated so far and computes a scalar p_{gen} denoting the probability of generating a new word from the vocabulary, $p_{gen} = \sigma(w_h^T h_t^* + w_s^T s_t + w_y^T y_t + b_{gen})$, where w_h, w_s, w_y, b_{gen} are trained vectors. The network probabilistically decides based on p_{gen}, whether to generate a new word from the vocabulary or copy a word from the source text using the attention distribution. For each word w in the vocabulary, the model calculates $P_{vocab}(w)$, the probability of the word getting generated next. For each word w' in the input article, its total attention received yields its probability of being copied. Since some words occur in the vocabulary and also the input article, they will have non-zero probabilities of being newly generated as well as being copied. Hence, the total probability of w being the next word generated in the summary(denoted by $\mathbf{p}$) is given by,

$$\mathbf{p}(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} a_i^t \tag{2}$$

The second term allows the framework to choose a word to copy from the input text using the attention distribution. The pointer-generator network further employs a coverage mechanism to encourage diversity in attention distributions over time steps. The training loss is set to be the average negative log-likelihood of the ground truth summaries. The model is trained using back-propagation and the Adagrad gradient descent algorithm [35]. Since the stylistic characteristics deal with specific expressions of the output text, it can be tailored by modifying $p(w)$ to incorporate the corresponding stylistic preferences. For more complex characteristics, as we show, it is possible to define a reinforcement learning based loss appended to the training loss to tailor the specific characteristics.

4 Content-specific Characteristics

Content-specific characteristics primarily govern what content needs to be presented in the output summary. For the rest of the paper, we illustrate the needs and modeling for content based characteristics with topical tailoring. However, the proposed approach can be extended to other content characteristics like entity-centric tailoring, etc.

Often, the whole content of the article may not be relevant to the readers and may prefer specific elements of the input to be summarized. For instance, a sports enthusiast may only be interested in content concerning that domain, or a surgeon may only be interested in health-related content. This calls for a need to generate multiple summary variants taking this information into account. Table 1 shows a particular instance from our dataset which talks about both Politics and Military. If a reader is interested only in Politics, the baseline summary generated by pointer-generator (**PGen**) model does not refer to politics and hence fails to meet the needs.

Article: bernie sanders , my vermont senator and , indeed , a friend of many years , is now running for president . he noted at his announcement -lrb- with a familiar note of wise irony -rrb- : " people should not underestimate me . " to most americans , of course , sen. bernie sanders is only a name , if that . he is barely known to the general public , which makes him a very long shot indeed to win election to the highest office in the nation . -lrb- cnn -rrb- he was impressively polite and bright in the eyes of his boyhood teachers , an encourager of his college friends . he was a docile captured killer in the care of paramedics tending to his gunshot wounds . dzhokhar " jahar " tsarnaev 's defense team is seeking to spare him from a death sentence for his part two years ago in the boston marathon bombings and murder of an mit police officer ...
PGen: dzhokhar " jahar " tsarnaev 's defense team is seeking to spare him from a death sentence . he was a docile captured killer in the care of paramedics tending to his gunshot wounds . tsarnaev was convicted april 8 on all 30 counts , including 17 that carry a possible death penalty .
Token-based (mixed) [5] + Attention Boosting; Politics: sen. bernie sanders is running for president . he is barely known to the general public , which makes him a very long shot indeed . he is barely known to the general public , which makes him a very long shot indeed .
Token-based (mixed) [5] + Attention Boosting; Military: dzhokhar " jahar " tsarnaev 's defense team is seeking to spare him from a death sentence . he was convicted april 8 on all 30 counts , including 17 that carry a possible death penalty . a paramedic testified wednesday that it was common for patients in shock to become agitated.

Table 1. Sample output topic-tailored summaries generated by Token-Based approach trained on CNN-DM-mixed dataset. We show just the top few sentences in the input article, in the interest of space.

Sequence-to-sequence learning models have been shown to understand where to look in the input through attention mechanisms which is then used for output generation. Tailoring content-specific characteristics would require this attention

to be tuned to focus on the relevant parts of the input (e.g. the relevant parts of the input talking about a topic of interest) to generate the desired output. This requires the model to be taught (either explicitly or implicitly) where to attend in the input article to tailor the summary appropriately.

One possibility is to maintain explicit indicators for each category of the characteristic (e.g. for each topic), allowing the model to learn where to pay more attention directly from the data. Fan et al. [4] propose to use such indicator tokens to tune characteristics such as length or desired entities. When training on an (article, summary) pair belonging to a particular bin, a token indicating the characteristic (topic in this case) represented by the summary is added to the beginning of the input article. While decoding an unseen article, the framework can generate multiple summary variants based on what token is prepended to the input word sequence. Internally, the model uses the token to learn a conditioned space of parameters, ensuring appropriate attention tuning to generate the summary with corresponding tailoring. We refer to this approach as **Token-Based** in our experiments.

A key requirement to make the model learn these intricacies is that the training data should contain sufficient samples under each category. However, there could be a skew in the dataset, which calls for alternate approach to tackle these characteristics. To deal with this problem, Krishna et al. [5] create a separate dataset where the model sees multiple summary variants for the same input article by mixing multi-topic articles. Given such a dataset, the vocabulary tokens can be used to guide the learning process towards a topic specific attention with a skewed-dataset. This method is called as **Token-Based (mixed)** which is trained on such an interspersed dataset.

While [4] and [5] use token based approaches implicitly teach the networks by taking advantage of the diversity in the training data, we propose an alternative to "explicitly" boost the attention distributions (referred to as **Attn-Boost**), restricting the model to focus on some parts of the input more than others. More formally, we modify Eq. 1 from the pointer generator as,

$$e_i^{t'} = \beta_i v^T \tanh(W_h h_i + W_s s_t + b_{att}), \tag{3}$$

where v, W_h, W_s and b_{att} are trainable model parameters as before and use this e_i to compute the attention a^t. β_i is a word specific attention boosting parameter. This explicitly teaches the model to pay more attention towards specific words than others. We leverage topic specific word lists curated by [5] and select the top k ($= 5$ in our experiments) sentences from the input article, which are most related to the target topic. We explicitly boost β_is of all the words in these sentences using the topic confidence measures as used by [5].

To draw more insights on each of these approaches, we evaluate it on the task of topic-based summarization on the CNN/DailyMail (**CNN-DM**) dataset [20, 19]. The dataset consists of $287,226$ training, $13,368$ validation and $11,490$ test instances. The articles have an average length of 781 tokens and multi-sentence summaries with average length of 56 tokens. We use the vanilla pointer generator (**PGen**) as the baseline for all our experiments retaining the hyper-parameters

by See et al. [3]. We train the **Token-Based** model [4] for topics by categorizing the ground-truth summaries into 9 topics: business, education, entertainment, health, military, politics, social, sports and technology (extending the setup by [5]) and prepending the topic to the input article while training.

Following [5], we also have a setup where we intersperse articles from different topics in **CNN-DM**, resulting in multiple topic-specific ground truth summaries for the same article. Using the 9 topics as before, the model now sees multiple summaries for the same article. There are $103,666$ article–summary-topic tuples for training, $4,720$ tuples for validation and $3,964$ tuples in the test dataset.

To evaluate the generation quality of all the approaches, we compare them on ROUGE 1, 2 and L F1-score. Note that generating summaries for all the topics may not make sense for the same input article, particularly when the article does not talk about the target topic. Hence, for each article in the test set, we generate the summary corresponding to the target topic defined by the ground truth summary. Then, we use the topic specific word lists [5] to get the top-1 and top-3 topics in the decoded summaries. The fraction of times the target topic lies in top-1 and top-3 topics of the decoded summaries, defines the **Top1** and **Top3** accuracies of the various setups.

Method	ROUGE F-score			% accuracy	
	1	**2**	**L**	**Top-1**	**Top-3**
PGen [3]	23.13	7.59	21.01	35.00	56.39
Token-Based [4]	26.53	9.05	23.98	37.52	63.36
Attn-Boost (Proposed)	22.05	7.30	19.95	40.06	62.85
Token-Based [4] + Attn-Boost	25.94	8.63	23.35	40.26	64.77
Token-Based (mixed) [5] + Attn-Boost	28.30	10.46	25.45	55.86	76.19

Table 2. Performance of proposed methodologies for generating topic-tuned summaries on the **CNN-DM-mixed** test dataset.

Table 2 summarizes the results of our methods for generating topic-oriented summaries. We observe that boosting attention values explicitly shows improvement in topic percentage accuracies but it suffers a decline in quality based on Rouge scores. This is expected since the explicit topic attention would make the model attend to parts of documents that are different from the ground truth summary. On the other hand, token based approach improves on ROUGE with a lesser topical accuracy. A combined framework of token-based and attention boosting yields the best performance across both ROUGE and topical accuracy metrics. This is perhaps because in the combined setup, the model learns more intricacies implicit in the data along with explicit attention to the topics - thus getting the best of both frameworks. When we train the same setup with the mixed dataset by [5], both ROUGE and the topical accuracies improve, suggesting the importance of the diversity in data for the network to implicitly learn attention patterns.

We show generated summaries from the token based approach in Table 1 for a particular instance from the testing dataset. The article was created by combining two articles from Politics and Military domain. The proposed approach appended with token-based framework on **CNN-DM-mixed** dataset is able to generate topic-specific variants, while the **PGen** approach fails to meet the re-

quirement towards both the topics. Figure 1 shows the average attention on the most attended parts of the input for the same instance, by the token based model trained on **CNN-DM-mixed**. When generating a 'Politics' oriented summary, the attention is on words like 'president', 'bernie sanders' and 'general public' showing the bias towards political phrases. On the other hand, when the target topic is 'Military', the focus/attention shifts to 'death sentence' and 'defence team'.
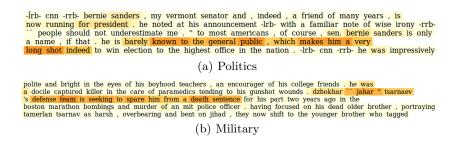
-lrb- cnn -rrb- bernie sanders , my vermont senator and , indeed , a friend of many years , is now running for president . he noted at his announcement -lrb- with a familiar note of wise irony -rrb- `` people should not underestimate me . '' to most americans , of course , sen. bernie sanders is only a name , if that . he is barely known to the general public , which makes him a very long shot indeed to win election to the highest office in the nation . -lrb- cnn -rrb- he was impressively

(a) Politics

polite and bright in the eyes of his boyhood teachers , an encourager of his college friends . he was a docile captured killer in the care of paramedics tending to his gunshot wounds . dzhokhar `` jahar '' tsarnaev 's defense team is seeking to spare him from a death sentence for his part two years ago in the boston marathon bombings and murder of an mit police officer . having focused on his dead older brother , portraying tamerlan tsarnav as harsh , overbearing and bent on jihad , they now shift to the younger brother who tagged

(b) Military

Fig. 1. Attention Distribution Over the Source Article for different Target Topics.

Our evaluations show that tuning content based characteristics can be achieved by modifying the attention of the network - either implicitly or explicitly. Implicit attention modification uses a token based approach [4], but relies on the diversity of the characteristics in the data. Where not available, feeding off of an interspersed dataset to artificially infuse diversity [5] is beneficial. By combining an interspersed dataset with explicitly attention boosting framework, the model is able to tune the characteristics better by learning where to attend and where-not to attend while tuning the content-based characteristics.

5 Stylistic Characteristics

Next, we focus on incorporating stylistic characteristics in the generated summary. Style-specific preferences ensure that the content is served appropriately to the target audience. In this direction, prior work has focused on incorporating dimensions such as sentiment [23], descriptiveness [26], formality [7] and many more across various tasks in text generation. We describe our methodology below to incorporate such characteristics into abstractive text summarization.

The way these stylistic aspects are incorporated into the sequence-to-sequence summarization framework, depends on how these aspects are defined. For example, simplicity of text can be defined at a lexical level based on the word-frequency in a simple corpus as defined in [36]. Krishna et al. [8] extend this towards generating simple summary by modifying the decoder probability in Eq. 2. They incorporate simplicity by modifying the beam search decoder to choose contextual replacement with words that are simpler; defining simplicity

as,

$$Simplicity(S) = \frac{1}{m} \sum_{i=1}^{m} \frac{f(s_i)}{1000}, \tag{4}$$

where $f(s_i)$ is the frequency of the ith word in the SUBTLEX [37] corpus. Krishna et al. [8] then use a word-to-word affinity/ replacement probabilities to achieve the tailoring.

However, not all aspects can be defined at a lexical level and hence, it is not always straightforward to modify the decoder probability. For example, readability can be quantified via the Flesch reading-ease score [38] given by,

$$206.835 - 1.015\frac{\text{total words}}{\text{total sentences}} - 84.6\frac{\text{total syllables}}{\text{total words}}. \tag{5}$$

The Flesch reading-ease score quantifies the difficulty in understanding a passage written in English. Higher scores indicate easier to read passages. It posits that the readability is inversely related to the average number of words in a sentence and to the average number of syllables in a word. Using a partial form of this definition, [8] propose to use shorter words (lesser syllables) as a surrogate to reading ease and use it to modify decoder probabilities. However, ignoring the first component of the reading ease makes this an incomplete tailoring.

Accounting for the first component would require the whole sentence to be generated before providing any feedback to the model on its generated readability. We propose to use a reinforcement learning based framework to incorporate such complex objectives requiring the generation of the complete/partial output for feedback generation. Recently, reinforcement learning frameworks have been successfully used to optimize text generation for content-based metrics, e.g. Rouge scores for summarization in [9], CIDEr scores for image captioning in [39]. We extend these to propose a reinforcement loss for stylistic elements as an additional term along with cross entropy using the Self Critical Sequence Training (SCST) [39] algorithm.

Given an input article word sequence x, and a corresponding ground truth summary $y^* = \{y_1^*, y_2^*, ..., y_T^*\}$, the pointer generator framework optimizes the negative log-likelihood objective function is given by,

$$L_{nll} = -\sum_{t=1}^{T} log\big[p(y_t^*|y_1^*, ..., y_{t-1}^*, x)\big]. \tag{6}$$

For providing explicit feedback on the stylistic characteristics, two output sequences are generated at the time of training: sampled sequence y^s and baseline sequence y^b. We generate y^s by sampling from the $p(y_t^s|y_1^s, y_2^s, .., y_{t-1}^s, x)$ distribution at each time step, and y^b by greedily choosing the word with maximum probability from the output distribution at each time step. The SCST algorithm defines a loss term L_{nll} with a reward for the target style characteristics,

$$L_{rl} = \big[r(y^b) - r(y^s)\big] \sum_{t=1}^{T} log\big[p(y_t^s|y_1^s, .., y_{t-1}^s, x)\big], \tag{7}$$

where $r(.)$ is the reward function - based on the target style characteristic to be optimized. Optimizing L_{rl} improves the expected reward of the generated output. The final loss is a linear combination of L_{nll} and L_{rl} given by,

$$L = (1 - \alpha) \cdot L_{nll} + (\alpha) \cdot L_{rl}, \tag{8}$$

where α governs the strength of RL based loss term. Reinforcement learning allows the loss function to include any non-differentiable metric in the form of rewards, which can be leveraged to optimize on our complex stylistic aspects directly. The Self Critical Sequence Training approach also helps in dealing with a exposure bias [9], a limitation in teacher forcing algorithm for training recurrent neural networks. By using sampled sequences, the model is exposed to its own distribution, learning to generate in accordance with such global meta properties.

To evaluate this methodology towards incorporating such characteristics, we use our setup to improve on the readability and simplicity of the generated summaries and incorporate the corresponding metrics into the learning algorithm directly as a reward function using the Reinforcement learning based loss function L_{rl}. To gain more insights on appropriate methods of tailoring stylistic characteristics, we compare our RL-based approach against the pointer-generator method from [3], the use of vocabulary tokens adapted from [4] and with modifying word-to-word affinity probabilities adapted from [8].

To adapt token based approach for readability [4], we define two tokens: "Not Readable" and "Readable" based on whether the readability of the ground truth summary was less or more as compared to the median value of 58.58 in the training dataset. We also evaluate against the lexical level modifications suggested by [8] and use their **VoTing** method to modify the generation probabilities by promoting the generation of shorter words over their longer synonyms. Finally, for our RL-based approach, we observe that training using the reinforcement loss is extremely slow, owing to the computation of sampled and greedy sequences along with the teacher forced outputs. Hence, we use the **PGen** model, pre-trained on **CNN-DM** dataset as the initialization point, i.e. training with $\alpha = 0$. Then, we train for $10,000$ more iterations with a fixed $\alpha = 0.9$.

Similarly, for simplicity, we leverage the work by [8] to measure simplicity based on Eq. 4 and establish the baselines similar to readability above. Our **RL-based** method directly uses this simplicity score as the reward function. For the token based approach, we divide the ground-truth summaries into two classes, "Not Simple" and "Simple" by thresholding at the median, observed to be 8.36 in the training dataset.

Similar to content-specific characteristics, we use ROUGE 1, ROUGE 2 and ROUGE L F-scores to evaluate the overlap between the generated and ground-truth summaries. To evaluate whether the models are able to capture our readability definition, we report the average Flesch reading-ease score for the generated summaries. For simplicity, we report the corresponding average simplicity score. Note that our objective is to understand whether the methods are able to capture a given definition for style-specific characteristics. Therefore, we have evaluated the tailoring based on the defined target metrics itself.

Table 3 summarizes our experiments for readability and simplicity. We observe a trade-off between the use of simpler/readable words from the vocabulary and the generation quality as captured by ROUGE metric primarily because of the deviation towards more simpler or readable words from the ones in reference summaries. The proposed RL-based approach is better able to capture readability, achieving higher average scores over all other approaches. However, it is not the best model for simplicity, where the lexical modifications at the decoder beats the RL method. This suggests that where the entire sequence needs to be generated to measure the stylistic aspect (like readability) it is useful to resort to RL-based frameworks. However, when the stylistic aspect can be measured lexically, decoder modifications perform better. There exist works in the reinforcement learning literature that have explored Actor-Critic methods [40] to provide intermediate feedback to the model even before generating the complete output sequences via partial rewards. Exploring such techniques to tackle simpler to complex definitions for style-specific constraints is a topic for future work.

Also note that token-based frameworks have a mixed results since it heavily relies on the diversity of training data without any explicit signal - hence might not be suited for stylistic aspects unless the training data contains sufficient diversity. It is possible to train a joint model which can be trained using the feedback on both ground-truth summaries in the data and the sequences sampled from the output distributions in a token based framework - which is a subject of further research.

Method	Readability				Simplicity			
	ROUGE F-score			Readability	ROUGE F-score			Simplicity
	1	2	L		1	2	L	
PGen [3]	**36.98**	**15.90**	**33.56**	53.17	**36.98**	**15.90**	**33.56**	9.37
Token-based [4]	36.84	15.85	33.56	55.35	35.70	15.17	32.50	9.58
VoTing [8]	36.91	15.72	33.48	54.10	33.42	13.03	30.3	**12.90**
RL-based	35.70	15.35	32.71	**57.50**	31.95	13.05	29.12	10.70

Table 3. Performance of the proposed approach in improving the readability and simplicity of generated summaries.

Table 4 shows the generated output summaries for the RL-based approach and **PGen** baseline model, on an instance from **CNN-DM** dataset, where our RL-based method achieves better readability scores using shorter sentence constructs. Such a framework can be used to teach the model on sentence level characteristics required to achieve a target style. Similarly the summaries generated by VoTing based approach is shown in Table 5 - the generated summaries uses simpler words in the summaries, such as 'big' in place of 'major' and 'hurt' in place of 'injured'. In a similar manner, careful modifications of the generated probabilities while decoding can be used to incorporate various other stylistic aspects defined at a lexical level.

Article(47.18): the killing of an employee at wayne community college in goldsboro , north carolina , may have been a hate crime , authorities said tuesday . investigators are looking into the possibility , said goldsboro police sgt. jeremy sutton . he did not explain what may have made it a hate crime . the victim – ron lane , whom officials said was a longtime employee and the school 's print shop operator – was white , as is the suspect . lane 's relatives said he was gay , cnn affiliate wncn reported . the suspect , kenneth morgan stancil iii , worked with lane as part of a work-study program , but was let go from the program in early march due to poor attendance , college president kay albertson said tuesday . on monday , stancil walked into the print shop on the third floor of a campus building , aimed a pistol-grip shotgun and fired once , killing lane , according to sutton . stancil has tattoos on his face ...
Reference(48.5): relatives of wayne community college shooting victim say he was gay , local media report . the suspect had worked for the victim but was let go , college president says . the suspect , kenneth morgan stancil iii , was found sleeping on a florida beach and arrested .
PGen(8.23): wayne community college , north carolina , may have been a hate crime , authorities say . investigators are looking into the possibility , said goldsboro police sgt. jeremy sutton . investigators are looking into the possibility , said goldsboro police sgt. jeremy sutton .
RL-based(50.12): the killing of an employee at wayne community college may have been a hate crime . the suspect , kenneth morgan stancil iii , worked with lane as part of a work-study program . he has no previous criminal record , authorities say .

Table 4. Sample output summary generated by incorporating readability as a reward function, along with baseline and reference summaries on an instance from **CNN-DM-mixed** dataset. The numbers in brackets refer to the corresponding readability scores. We show just the top few sentences in the input article, in the interest of space.

6 Conclusions

In this work, we study a variety of constraints which may be imposed while generating abstractive summaries of a given input article by categorizing these constraints as either content-specific, which govern what content needs to be generated and style-specific, which govern various stylistic expressions in these outputs. Our experiments indicate that the content-based characteristics can be tailored in the summary via explicitly or implicitly tuning the attention to focus on relevant parts of the network. Approach to tailor stylistic constraints depends on the nature of definition - characteristics defined at lexical level can be tuned better by modifying decoder probabilities during beam search. More complicated metrics can be tuned by using reinforced rewards in the loss function.

References

1. Nenkova, A., McKeown, K.: Automatic summarization. Foundations and Trends® in Information Retrieval (2011)

Article: hong kong -lrb- cnn -rrb- six people were hurt after an explosion at a controversial chemical plant in china 's southeastern fujian province sparked a huge fire , provincial authorities told state media . the plant , located in zhangzhou city , produces paraxylene -lrb- px -rrb- , a reportedly carcinogenic chemical used in the production of polyester films and fabrics . the blast occurred at an oil storage facility monday night after an oil leak , though local media has not reported any toxic chemical spill ...

Summary: ...five out of six people were **hurt** *(injured)* by broken glass and have been sent to the hospital for treatment .

Article: -lrb- cnn -rrb- debates on climate change can break down fairly fast . there are those who believe that mankind 's activities are changing the planet 's climate , and those who do n't . but a new way to talk about climate change is emerging , which shifts focus from impersonal discussions about greenhouse gas emissions and power plants to a very personal one : your health ...

Summary: ...it 's easy to brush aside debates involving **big** *(major)* international corporations , but who would n't stop to think and perhaps do something about their own health

Table 5. Sample simplified summaries generated by the proposed approach. Words in bold show the use of simpler summaries generated by our approach, while the words in italics are those picked up by the baseline model.

2. Nallapati, R., Zhai, F., Zhou, B.: Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In: AAAI. (2017) 3075–3081
3. See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. In: Annual Meeting of the Association for Computational Linguistics (ACL). (2017)
4. Fan, A., Grangier, D., Auli, M.: Controllable abstractive summarization. arXiv preprint arXiv:1711.05217 (2017)
5. Krishna, K., Srinivasan, B.V.: Generating topic-oriented summaries using neural attention. In: Conference of the North American Chapter of the Association for Computational Linguistics(NAACL). (2018)
6. Wang, L., Yao, J., Tao, Y., Zhong, L., Liu, W., Du, Q.: A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. arXiv preprint arXiv:1805.03616 (2018)
7. Niu, X., Martindale, M., Carpuat, M.: A study of style in machine translation: Controlling the formality of machine translation output. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. (2017) 2814–2819
8. Krishna, K., Murhekar, A., Sharma, S., Srinivasan, B.V.: Vocabulary tailored summary generation. In: International Conference on Computational Linguistics (COLING). (2018)
9. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. arXiv (2018)
10. Wang, L., Cardie, C.: Domain-independent abstract generation for focused meeting summarization. In: ACL (1). (2013) 1395–1405
11. Genest, P.E., Lapalme, G.: Framework for abstractive summarization using text-to-text generation. In: Workshop on Monolingual Text-To-Text Generation. (2011)

12. Filippova, K.: Multi-sentence compression: Finding shortest paths in word graphs. In: International Conference on Computational Linguistics (COLING). (2010)
13. Berg-Kirkpatrick, T., Gillick, D., Klein, D.: Jointly learning to extract and compress. In: Annual Meeting of the Association for Computational Linguistics (ACL). (2011)
14. Banerjee, S., Mitra, P., Sugiyama, K.: Multi-document abstractive summarization using ilp based multi-sentence compression. In: IJCAI. (2015) 1208–1214
15. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. (2014) 3104–3112
16. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. In: Conference on Empirical Methods in Natural Language Processing. (2015)
17. Chopra, S., Auli, M., Rush, A.M., Harvard, S.: Abstractive sentence summarization with attentive recurrent neural networks. In: HLT-NAACL. (2016) 93–98
18. Takase, S., Suzuki, J., Okazaki, N., Hirao, T., Nagata, M.: Neural headline generation on abstract meaning representation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. (2016) 1054–1059
19. Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., Xiang, B.: Abstractive text summarization using sequence-to-sequence rnns and beyond. In: SIGNLL Conference on Computational Natural Language Learning. (2016)
20. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems. (2015) 1693–1701
21. Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., Bengio, Y.: Pointing the unknown words. In: Annual Meeting of the Association for Computational Linguistics (ACL). (2016)
22. Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H.: Modeling coverage for neural machine translation. In: 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). (2016)
23. Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., Xing, E.P.: Toward controlled generation of text. In: International Conference on Machine Learning. (2017) 1587–1596
24. Shen, T., Lei, T., Barzilay, R., Jaakkola, T.: Style transfer from non-parallel text by cross-alignment. In: Advances in Neural Information Processing Systems. (2017) 6833–6844
25. Sennrich, R., Haddow, B., Birch, A.: Controlling politeness in neural machine translation via side constraints. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (2016) 35–40
26. Ficler, J., Goldberg, Y.: Controlling linguistic style aspects in neural language generation. In: Proceedings of the Workshop on Stylistic Variation. (2017) 94–104
27. Niu, X., Rao, S., Carpuat, M.: Multi-task neural models for translating between styles within and across languages. arXiv preprint arXiv:1806.04357 (2018)
28. Oraby, S., Reed, L., Tandon, S., Sharath, T., Lukin, S., Walker, M.: Controlling personality-based stylistic variation with neural natural language generators. arXiv preprint arXiv:1805.08352 (2018)
29. Tikhonov, A., Yamshchikov, I.P.: What is wrong with style transfer for texts? arXiv preprint arXiv:1808.04365 (2018)
30. Artetxe, M., Labaka, G., Agirre, E., Cho, K.: Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041 (2017)

31. Han, M., Wu, O., Niu, Z.: Unsupervised automatic text style transfer using lstm. In: National CCF Conference on Natural Language Processing and Chinese Computing, Springer (2017) 281–292
32. Xu, J., Sun, X., Zeng, Q., Ren, X., Zhang, X., Wang, H., Li, W.: Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. arXiv preprint arXiv:1805.05181 (2018)
33. Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., Black, A.W.: Style transfer through back-translation. arXiv preprint arXiv:1804.09000 (2018)
34. Zhang, Y., Ding, N., Soricut, R.: Shaped: Shared-private encoder-decoder for text style adaptation. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Volume 1. (2018) 1528–1538
35. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research (2011)
36. Paetzold, G., Specia, L.: Lexenstein: A framework for lexical simplification. Proceedings of ACL-IJCNLP 2015 System Demonstrations (2015) 85–90
37. Brysbaert, M., New, B.: Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. Behavior research methods **41** (2009) 977–990
38. Flesch, R.F.: How to write plain English: A book for lawyers and consumers. Harpercollins (1979)
39. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR. (2017)
40. Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., Bengio, Y.: An actor-critic algorithm for sequence prediction. arXiv preprint arXiv:1607.07086 (2016)