Inference Time Style Control for Summarization

Shuyang Cao and **Lu Wang**
Computer Science and Engineering
University of Michigan
Ann Arbor, MI
{caoshuy, wangluxy}@umich.edu

Abstract

How to generate summaries of different styles without requiring corpora in the target styles, or training separate models? We present two novel methods that can be deployed during summary decoding on any pre-trained Transformer-based summarization model. (1) *Decoder state adjustment* instantly modifies decoder final states with externally trained style scorers, to iteratively refine the output against a target style. (2) *Word unit prediction* constrains the word usage to impose strong lexical control during generation. In experiments of summarizing with simplicity control, automatic evaluation and human judges both find our models producing outputs in simpler languages while still informative. We also generate news headlines with various ideological leanings, which can be distinguished by humans with a reasonable probability.

1 Introduction

Generating summaries with different language styles can benefit readers of varying literacy levels (Chandrasekaran et al., 2020) or interests (Jin et al., 2020). Significant progress has been made in abstractive summarization with large pre-trained Transformers (Dong et al., 2019; Lewis et al., 2020; Zhang et al., 2019; Raffel et al., 2019; Song et al., 2019). However, style-controlled summarization is much less studied (Chandrasekaran et al., 2020), and two key challenges have been identified: (1) *lack of parallel data*, and (2) *expensive (re)training*, e.g., separate summarizers must be trained or fine-tuned for a pre-defined set of styles (Zhang et al., 2018). Both challenges call for inference time methods built upon trained summarization models, to adjust styles flexibly and efficiently.

To address these challenges, we investigate *just-in-time style control techniques that can be directly applied to any pre-trained sequence-to-sequence (seq2seq) summarization model*. We study two methods that leverage external classifiers to favor the generation of words for a given style. First, **decoder state adjustment** is proposed to alter the decoder final states with feedback signaled by style scorers, which are trained to capture global property. Second, to offer stronger *lexical control*, we introduce **word unit prediction** that directly constrains the output vocabulary. Example system outputs are displayed in Fig. 1. Notably, our techniques are deployed at *inference time* so that the summary style can be adaptively adjusted during decoding.

Daily Mail Article: . . . [A 16-year-old who was born a girl but identifies as a boy has been **granted the opportunity** to go through male puberty thanks to hormone treatment.] . . . [The transgender boy, who has felt as though he is living in the wrong body since he was a child, has been given permission by a Brisbane-based judge to receive testosterone injections] . . .

(a) Decoder State Adjustment: [Queensland teen has been granted hormone treatment. The 16-year-old was born a girl but identifies as a boy.] . . . [A judge has granted the teen permission to receive testosterone injections.] . . .
(b) Word Unit Prediction: A 16-year-old who was born a girl has been **given the right** to go through male puberty. The transgender boy has lived in a female body since he was a . . .

Figure 1: Sample summaries generated by our style control methods via (a) adjusting decoder states with a simplicity scorer and (b) predicting simple words to use. **Gray** texts are produced by BART but removed after decoder state adjustment. Simplified words and their counterparts in the source are highlighted in **blue**.

We experiment with two tasks: (1) simplicity control for document summarization with CNN/Daily Mail, and (2) headline generation with various ideological stances on news articles from the SemEval task (Kiesel et al., 2019) and a newly curated corpus consisting of multi-perspective stories from AllSides[1]. In this work, the algorithms are experimented with the BART model (Lewis et al., 2020), though they also work with other Transformer models. Both automatic and human

[1] www.allsides.com

evaluations show that our models produce summaries in simpler languages than competitive baselines, and the informativeness is on par with a vanilla BART. Moreover, headlines generated by our models embody stronger ideological leaning than nontrivial comparisons.[2]

2 Related Work

Summarizing documents into different styles are mainly studied on news articles, where one appends style codes as extra embeddings to the encoder (Fan et al., 2018), or connects separate decoders with a shared encoder (Zhang et al., 2018). Similar to our work, Jin et al. (2020) leverage large pre-trained seq2seq models, but they modify model architecture by adding extra style-specific parameters. Nonetheless, existing work requires training *new* summarizers for different target styles or modifying the model structure. In contrast, our methods only affect decoder states or lexical choices during inference, allowing on-demand style adjustment for summary generation.

Style-controlled text generation has received significant research attentions, especially where parallel data is scant (Lample et al., 2019; Shang et al., 2019; He et al., 2020). Typical solutions involve disentangling style representation from content representation, and are often built upon autoencoders (Hu et al., 2017) with adversarial training objectives (Yang et al., 2018). The target style is then plugged in during generation. Recently, Dathathri et al. (2020) propose plug and play language models (PPLMs) to alter the generation style by modifying all key-value pairs in the Transformer, which requires heavy computation during inference. Krause et al. (2020) then employ a generative discriminator (GeDi) to improve efficiency. Our methods are more efficient since we only modify the decoder final states or curtail the vocabulary.

3 Inference Time Style Control

3.1 Global Characteristic Control via Decoder State Adjustment

Given a style classifier $q(z|\cdot)$ that measures to which extent does the current generated summary resemble the style z, we use its estimate to adjust the final decoder layer's state $\mathbf{o}_t$ at step t with gradient descent, as illustrated in Fig. 2. The

[2]Our code and data are available at: https://shuyangcao.github.io/projects/inference_style_control.

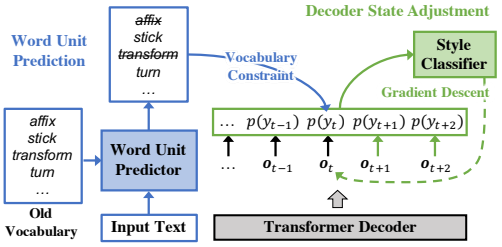

Figure 2: Just-in-time style control: (1) Decoder state adjustment takes in a style score and iteratively updates $\mathbf{o}_t$; (2) Word unit prediction controls the vocabulary.

output token is produced as $p(y_t|y_{1:t-1}, \mathbf{x}) = \text{softmax}(\mathbf{W}_e\mathbf{o}_t)$, $\mathbf{W}_e$ is the embedding matrix.

Concretely, to generate the t-th token, a style score of $q(z|\mathbf{y}_{1:t+2})$ is first computed. In addition to what have been generated up to step $t-1$, we also sample y_t and two future tokens for style estimation. The decoder state is updated as follows:

$$\mathbf{o}_t \leftarrow \mathbf{o}_t - \lambda \nabla_{\mathbf{o}_t} \big[-q(z|\mathbf{y}_{1:t+2}) \big] \quad (1)$$

where λ is the step size. Gradient descent is run for 10 iterations for document summarization and 30 iterations for headline generation.

Below, we define one discriminative and one generative style classifier, to illustrate the method.

Discriminative Style Scorer. We feed the tokens into a RoBERTa encoder (Liu et al., 2019) and use the contextualized representation of the `BOS` token, i.e., $\mathbf{h}_0$, to predict the style score as $p_{sty}(z|\cdot) = \text{softmax}(\mathbf{W}_s\mathbf{h}_0)$, where $\mathbf{W}_*$ are learnable parameters in this paper. At step t of summary decoding, the style score is estimated as:

$$q(z|\mathbf{y}_{1:t+2}) = \log p_{sty}(z|\mathbf{y}_{1:t+2}) \quad (2)$$

For the discriminative style scorer, the step size λ is set to 1.0.

Generative Language Model Scorer. We build a class-conditional language model (CC-LM) from texts prepended with special style-indicating tokens. Concretely, the CC-LM yields probabilities $p_{LM}(y_{t'}|\mathbf{y}_{1:t'-1}, z)$ ($p_{LM}(y_{t'}, z)$ for short), conditional on the previously generated tokens $\mathbf{y}_{1:t'-1}$ and the style z. As the summarizer's output probability $p(y_{t'})$ should be close to the language model's estimate, the style score is defined as:

$$q(z|\mathbf{y}_{1:t+2}) = \frac{1}{t+2} \sum_{t'=1}^{t+2} p_{LM}(y_{t'}, z) \log p(y_{t'}) \quad (3)$$

Here we use a step size λ of 0.1.

3.2 Lexical Control via Word Unit Prediction

Lexical control is another tool for managing summary style, as word choice provides a strong signal of language style. Given an input document, our goal is to predict a set of word units (e.g., the subwords used in BART pre-training) that can be used for summary generation. For instance, if the input contains "affix", we will predict "stick" to be used, while excluding the original word "affix". A similar idea has been used to expedite sequence generation (Hashimoto and Tsuruoka, 2019), though our goal here is to calculate the possibilities of different lexical choices.

Concretely, after encoding the input $\mathbf{x}$ by RoBERTa, we take the average of all tokens' contextual representations, and pass it through a residual block (He et al., 2016) to get its final representation $\tilde{\mathbf{R}}$. We then compute a probability vector for all word units in the vocabulary as $\mathbf{p}^r = \text{sigmoid}(\mathbf{W}_r\tilde{\mathbf{R}})$. The top v word units with the highest probabilities are selected and combined with entity names from the input to form the new vocabulary, from which the summary is generated. We use $v = 1000$ in all experiments.

Dynamic Prediction. We also experiment with a dynamic version, where the word unit predictor further considers what have been generated up to a given step. In this way, the new vocabulary is updated every m steps ($m = 5$ for document summarization, and $m = 3$ for headline generation).

4 Simplicity-controlled Document Summarization

For experiments, we use BART fine-tuned on the CNN/DailyMail (CNN/DM) (Hermann et al., 2015), by following Lewis et al. (2020) for data preprocessing and splitting. The numbers of data in train, validation and test splits are 287,188, 13,367 and 11,490, respectively.

We use paragraph pairs from normal and simple English Wikipedia articles in Hua and Wang (2019) for *simplicity style scorer* and *class-conditional language model* training. We split the pairs into 86,467, 10,778, and 10,788 for training, validation and testing, respectively. On the test set, our simplicity style scorer achieves an F1 score of 89.7 and our class-conditional language model achieves a perplexity of 30.35.

To learn the *word unit predictor*, for each paragraph pair, the predictor reads in the normal version and is trained to predict the word units used in the simple version. For the *dynamic version*, it predicts which word units are used to generate the rest of the text, after every 5 steps. Recalls for the two predictors on the test set are 81.5 and 80.0.

Model	Style			Flu.	Cont.
	Simp.↑	%Simp.↑	Rd.↓	PPL↓	BERT↑
BART	56.93	62.70	8.06	34.05	88.62
RERANKING	71.33	62.68	8.04	36.17	88.62
LBLCTRL	56.21	62.71	8.07	28.85	88.57
CTRLGEN	81.56	64.78	7.79	70.36	88.01
TRANS	59.78	63.03	7.99	33.17	88.46
GEDI	71.33	62.57	7.88	33.48	**88.79**
LIGHTLS	69.02	64.92	7.72	76.37	86.98
Ours w/ Decoder State Adjustment					
SIMP. SCORER	86.67	62.94	7.77	34.20	88.71
SIMP. CC-LM	75.04	64.27	7.69	30.49	88.73
Ours w/ Word Unit Prediction					
WORDU	**95.85**	67.23	**7.19**	**27.40**	87.76
DYNAMIC WORDU	93.87	**67.37**	7.23	28.42	87.91

Table 1: Automatic evaluation on summarization with simplicity, with simplicity level by our scorer (Simp., probability multiplied by 100), % of words in the Dale-Chall simple word list (%Simp.), Dale-Chall readability (Rd.), fluency by perplexity (PPL), and content metric by BERTScore (BERT). Our models are significantly better than the comparisons ($p < 0.005$) on simplicity and readability, except for CTRLGEN and LIGHTLS.

For comparison, we consider **RERANKING** beams based on our style score at the last step. We also use a label-controlled (**LBLCTRL**) baseline as described in Niu and Bansal (2018), where summaries in the training data are labeled as simple or normal by our scorer. We further compare with **GEDI** and two pipeline models: a style transfer model (Hu et al., 2017) applied on the output of BART (**CTRLGEN**) and a normal-to-simple translation model fine-tuned from BART (**TRANS**), both trained on Wikipedia. Finally, we consider **LIGHTLS** (Glavaš and Štajner, 2015), a rule-based lexical simplification model.

Automatic Evaluation. Table 1 shows that *our models' outputs have significantly better simplicity and readability while preserving fluency and a comparable amount of salient content*. Key metrics include simplicity level estimated by our scorer and Dale-Chall readability (Chall and Dale, 1995). We use GPT-2 perplexity (Radford et al., 2019) to measure fluency, and BERTScore (Zhang* et al., 2020) for content preservation. Our inference time style control modules can adaptively change the output style, and thus outperform reranking at the end of generation or using pipeline models. More-

Model	Inf.↑	Flu.↑	Simp.R.↓	Top 1↑
BART	4.45	**4.90**	2.19	19.0%
GEDI	4.48	4.83	2.00	23.8%
SIMP. SCORER	**4.53**	4.83	1.66*	48.4%
DYNAMIC WORDU	4.36	4.84	**1.65***	**57.9%**

Table 2: Human evaluation on informativeness (Inf.), fluency (Flu.), simplicity ranking (Simp.R.), and percentage of summaries ranked as simplest (Top 1). Krippendorff's α: 0.38, 0.22, and 0.16 (first three metrics). $*$: significantly better than comparisons ($p < 0.005$).

over, by iteratively adjusting the decoder states, our methods deliver stronger style control than GEDI, which only adjusts the probability once per step.

When comparing among our models, we find that *word unit prediction is more effective at lexical simplification than updating decoder states*, as demonstrated by the higher usage of simple words according to the Dale-Chall list. We believe that strong lexical control is achieved by directly pruning output vocabulary, whilst decoder state adjustment is more poised to capture global property, e.g., sentence compression as shown in Fig. 1. Moreover, we compute the edit distance between our style-controlled system outputs and the summaries produced by the fine-tuned BART. We find that adjusting decoder states with style scorer and language model yields an edit distance of 45.7 and 47.4, compared to larger distances of 56.7 and 54.3 given by word unit prediction and with additional dynamic prediction.

Human Evaluation. We recruit three fluent English speakers to evaluate system summaries for **informativeness**—whether the summary covers important information from the input, and **fluency**—whether the summary is grammatical, on a scale of 1 (worst) to 5 (best). They then rank the summaries by **simplicity** level (ties are allowed). 50 samples are randomly selected for evaluation, and system summaries are shuffled. As seen in Table 2, *summaries by our models are considered simpler* than outputs of BART and GEDI, with better or comparable informativeness.

5 Ideology-controlled Headline Generation

To generate news headlines of various ideological leanings, we use the **SemEval** Hyperpartisan News Detection dataset (Kiesel et al., 2019), where each article is labeled with a stance: *left*, *leaning left*, *neutral*, *leaning right*, or *right*. Here, we combine left and leaning-left articles into one bucket, and

Model	Left		Right	
	Ideol.	BERT	Ideol.	BERT
BART	18.63	91.03	19.04	91.03
RERANKING	30.80	90.68	30.11	90.66
LBLCTRL	20.59	**90.97**	20.89	**91.02**
GEDI	12.64	84.84	3.61	84.84
Ours w/ Decoder State Adjustment				
IDEOL. SCORER	**31.15**	90.08	**30.54**	90.17
IDEOL. CC-LM	23.74	89.65	20.79	89.65
Ours w/ Word Unit Prediction				
WORDU	21.30	89.64	20.42	90.13
DYNAMIC WORDU	21.53	89.49	20.09	90.19

Table 3: Ideological headline generation results. Using ideology scorer to update decoder states yields the highest ideology scores (multiplied by 100).

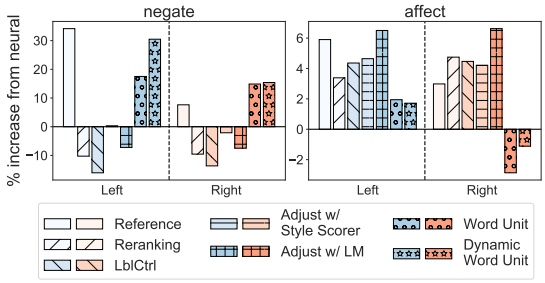

Figure 3: LIWC word usage changes of "negate" and "affect", compared to neutral headlines. In each subfigure, left and right panels correspond to left and right leaning stances.

similarly for right and leaning-right articles. We use the lead paragraph as the input, and the headline as the target generation. The data is processed following Rush et al. (2015), and split into 346,985 for training, 30,000 each for validation and testing. Details of the ideology distribution for SemEval are in Appendix B.

We fine-tune BART and train ideology classifiers on the SemEval training set. First, two binary *style scorers* are trained on headlines of left and right stances, with F1 scores of 76.1 and 78.0, respectively. One *class-conditional language model* is trained on headlines with a stance token (left or right) prepended, achieving a perplexity of 54.7. To learn the *word unit predictor* for the left (and similarly for the right), we use samples that are labeled as left-leaning, treat the lead paragraph as the input, and then predict the word units used in the headline. Recalls for our predictors range from 77.8 to 83.5.

Automatic Evaluation with SemEval. Table 3 shows that *our decoder state adjustment model with the ideology scorer obtains the highest ideology scores*, due to its effectiveness at capturing

Model	Rel.	Edit	Hmn	Hmn Acc.
Human	4.01	12.24	60.8%	73.3%
RERANKING	**4.71**	3.90	24.5%	52.5%
LBLCTRL	4.70	2.30	11.6%	**71.4%**
IDEOL. SCORER (ours)	4.47	**8.86***	**42.5%***	53.9%
DYNAMIC WORDU (ours)	4.66	4.20	25.8%	51.6%

Table 4: Human evaluation of ideology-controlled headline generation with relevance (Rel.), edit distance (Edit) between left and right headlines, % of samples perceived as having different stances (Hmn), and (among them) accuracy of identified stances (Hmn Acc.). Krippendorff's α of relevance: 0.48. $*$: significantly better than other models ($p < 0.005$).

Paragraph: The Obama administration on Thursday rolled out new efforts aimed at curtailing gun violence . . .
REFERENCE
[L]: obama offers new executive actions on gun control
[R]: administration announces new gun control measures, targets military surplus imports
IDEOL.SCORER
[L]: u.s. moves to **curb gun violence** with new rules
[R]: obama admin to **tighten gun control laws**
DYNAMIC WORDU
[L]: obama unveils new steps to **curb gun violence**
[R]: obama administration unveils new **gun control** measures

Table 5: Sample generated headlines with left (shaded in blue) and right (red) stances. Phrases that are typically used by a stance are in **bold**.

the global context—stance is often signaled by the joint selection of entities and sentiments.

One might be interested in *which words are favored for ideology-controlled generation*. To that end, we analyze the change of word usages with Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015). In Fig. 3, it can be seen that word unit prediction-based models generate more "negations", consistent with trends observed in human-written headlines. Meanwhile, models with decoder state adjustment and the baselines all use more "affect" words in both stances, indicating that they consider it easier to use explicit sentiments to demonstrate the stances.

Human Evaluation with AllSides. Given the low ideology scores in Table 3, we further study *if human can distinguish the stances in human-written and system generated headlines*. News clusters from **AllSides** are used, where each cluster focuses on one story, with multiple paragraph-headline pairs from publishers of *left*, *neutral*, and *right* ideological leanings. We use the lead paragraph as the input, and collect 2,985 clusters with samples written in all three stances. More details of the collection are in Appendix B. We test and report results by using lead paragraphs from *neural* articles as the input to construct headlines of *left* and *right* ideological stances.

We randomly pick 80 samples and include, for each sample, two headlines of different stances generated by each system. Raters first score the **relevance** of the generated headlines to the neutral paragraph's headline, on a scale of 1 to 5. They then read each pair of headlines to decide whether they are written in different stances, and if so, to label them. Table 4 highlights the intrinsic difficulty of capturing ideological language usage: Even reference headlines are only distinguishable in 60.8% of the cases, among which the stance identification accuracy is 73.3%. In comparison, 42.5% of the output pairs by the decoder state adjustment model can be distinguished, significantly higher than those of the baselines (24.5% and 11.6%). Sample outputs by our models are shown in Table 5, with more outputs included in Appendix E.

6 Conclusion

We present two just-in-time style control methods, which can be used in any Transformer-based summarization models. The decoder state adjustment technique modifies decoder final states based on externally trained style scorers. To gain stronger lexical control, word unit prediction directly narrows the vocabulary for generation. Human judges rate our system summaries to be simpler with better readability. We are also able to generate headlines with different ideological leanings.

Acknowledgements

This research is supported in part by National Science Foundation through Grant IIS-1813341, and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. We thank all the anonymous reviewers for their constructive suggestions.

References

J.S. Chall and E. Dale. 1995. *Readability revisited: the new Dale-Chall readability formula*. Brookline Books.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview of the first workshop on scholarly document processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13063–13075. Curran Associates, Inc.

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

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2019. Accelerated reinforcement learning for sentence generation by vocabulary prediction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3115–3125, Minneapolis, Minnesota. Association for Computational Linguistics.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 591–602, Hong Kong, China. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5093, Online. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *12th International Workshop on Semantic Evaluation (SemEval 2019)*. Association for Computational Linguistics.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report, Austin, TX: University of Texas at Austin.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

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

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4937–4946, Hong Kong, China. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936, Long Beach, California, USA. PMLR.

Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7287–7298. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

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

Ye Zhang, Nan Ding, and Radu Soricut. 2018. 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)*, pages 1528–1538, New Orleans, Louisiana. Association for Computational Linguistics.

A Training and Decoding Settings

Training. We train our *simplicity style scorer* and *ideology style scorers* for 10 epochs. The peak learning rate is 1×10^{-5} with a batch size of 32.

The *class-conditional language models* for simplicity and ideology are trained with a peak learning rate of 5×10^{-4} until the perplexity stops dropping on the validation set. We limit the number of tokens in each batch to $2,048$.

All *word unit predictors* are trained with a peak learning rate of 1×10^{-4} until the loss on the validation set no longer drops. We use a batch size of 32 for training.

Decoding. We use beam search for decoding. A beam size of 5 is used for all models except for the decoder state adjustment having a beam size 1 (greedy decoding) to maintain a reasonable running time. Repeated trigrams are disabled for generation in all experiments. As suggested by Lewis et al. (2020) and Yan et al. (2020), length penalties are set to 2.0 and 1.0 for summarization and headline generation, respectively. The minimum and maximum lengths are set for decoding at 55 and 140 for summarization, 0 and 75 for headline generation.

B Statistics on SemEval and Allsides

Each article in the SemEval dataset is labeled with a stance: *left*, *leaning left*, *neutral*, *leaning right*, or *right*. Here we combine left and leaning-left

Split	Left	Neutral	Right
Training	122,449	86,472	138,064
Validation	10,000	10,000	10,000
Test	10,000	10,000	10,000

Table 6: Ideology distribution for training, validation and test set splits of SemEval.

articles into one bucket, and similarly for right and leaning-right articles. The ideology distribution for training, validation and test splits are in Table 6.

In our human evaluation of ideology-controlled headline generation, we use data collected from Allsides. The Allsides news clusters are curated by editors. The stance labels for different publishers are provided by Allsides, which are synthesized from blind surveys, editorial reviews, third-party analyses, independent reviews, and community feedback. We collect all the Allsides news clusters by April 26, 2020. After removing empty clusters, the total number of news clusters is 4,422. Among them, 2,985 clusters contain articles written in all three stances. For each article in the cluster, we keep the first paragraph and pair it with the headline. We remove the bylines in the first paragraphs.

C Additional Results for Headline Generation

In Table 7, we show the results of ideology-controlled headline generation on SemEval with BART fine-tuned on Gigaword (Napoles et al., 2012). Our methods are still effective, especially by using decoder states adjustment with style scorers.

Model	Left		Right	
	Ideol.	BERT	Ideol.	BERT
BART	21.77	88.81	20.72	88.81
Ours w/ Decoder State Adjustment				
IDEOL. SCORER	**39.61**	87.96	**34.14**	87.89
IDEOL. CC-LM	27.38	87.79	22.21	87.76
Ours w/ Word Unit Prediction				
WORDU	22.98	88.35	21.09	88.40
DYNAMIC WORDU	22.84	88.32	21.08	88.47

Table 7: Ideological headline generation results with BART fine-tuned on the Gigaword dataset.

D Human Evaluation Guidelines

We include the evaluation guidelines for summarization and headline generation in Figures 4 and 5.

E Sample Outputs

Additional outputs are in Figures 6 and 7.

	Article
	There was no special treatment for Lewis Ferguson at Paul Nicholls's yard on Thursday morning. The 18-year-old was mucking out the stables as usual, just a cut on the nose to show for the fall which has made him an internet sensation. Ferguson's spectacular double somersault fall from the favourite Merrion Square in the 4.20pm at Wincanton has been watched hundreds of thousands of times online. But he was back riding out and is undeterred from getting back in the saddle. Amateur jockey Lee Lewis Ferguson has just a cut on his nose to show for his ordeal . Teenager Ferguson was flung from his horse in spectacular fashion at Wincanton . 'It was just a blur,' he said. 'I couldn't work out what had happened until I got back to the weighing room and watched the replay. All the other jockeys asked me if I was all right and stuff, they all watched with me and looked away in horror. (....)
	Informativeness:
1	Not relevant to the article
	e.g., *"Paul Nicholl's yard will start its expansion in December. The expansion plan was carried out six months ago."*
3	Relevant, but misses the main point
	e.g., *"Amateur jockey Lee Lewis Ferguson has just a cut on his nose to show for his ordeal . 'It was just a blur,' he said."*
5	Successfully captures the main point and most of the important points.
	e.g., *"Lewis Ferguson was mucking out the stables as usual on Thursday. Favourite Merrion Square threw jockey in a freak fall on Wednesday."*
	Fluency:
1	Summary is full of garbage fragments and is hard to understand
	e.g., *"18 year old nose. to cut show nose. the horse fashion, as to"*
2	Summary contains fragments, missing components but has some fluent segments
	e.g., *"Lewis Ferguson out on Thursday. threw jockey on Wednesday."*
3	Summary contains some grammar errors but is in general fluent
	e.g., *"Lewis Ferguson was muck out the stables as usual onThursday. The Merrion Square threw jockey jockey in a freak fall on Wednesday. His spectacular doublesomersault fall made him internetsensation."*
4	Summary has relatively minor grammatical errors
	e.g., *"Lewis Ferguson was mucking out the stables as usual on in Thursday. Favourite Merrion Square threw jockey ina freak fall on Wednesday. His spectacular double somersault fall made him internet sensation."*
5	Fluent Summary
	e.g., *"Lewis Ferguson was mucking out the stables as usual on Thursday. Favourite Merrion Square threw jockey in a freak fall on Wednesday. His spectacular double somersault fall made him internet sensation."*
	Simplicity:
Bad	The summary uses complex words that can be replaced with simpler ones in almost all sentences and complex syntax structures (e.g., two or more clauses in a sentence)
	e.g., *"Lewis Ferguson was thrown by Merrion Square and made a spectacular double somersault fall which gathered millions of views online, making him internet sensation. But he was back riding out and is undeterred from getting back in the saddle, just a cut on the nose to show for the fall ."*
Moderate	The summary uses at most one complex words that can be replaced with simpler ones per sentence, and uses syntax structures with at most one clause in a sentence
	e.g., *"Lewis Ferguson fell from Merrion Square. His spectacular double somersault fall made him internet sensation. But he was back riding out and is not afraid of getting back in the saddle."*
Good	The summary almost always uses simple and common words and simple syntax structures (e.g., no clause or at most one clause in the whole summary)
	e.g., *"Lewis Ferguson fell from his horse on Wednesday. His eye-catching double flip fall made him famous on the Internet. He was back to the yard. He is not afraid of getting back in the saddle."*

Figure 4: Sample summaries with explanations on human evaluation aspect scales and examples of summaries at different simplicity levels.

	Paragraph
	US President Donald Trump has said he is going to halt funding to the World Health Organization (WHO) because it has "failed in its basic duty" in its response to the coronavirus outbreak.
	Relevance:
1	The headline does not contain any information related to the input
	e.g., *"'a hateful act': what we know about the ft. lauderdale airport shooting"*
2	The headline contains some relevant event or person in the paragraph, but the topic is largely irrelevant
	e.g., *"trump: i don't take questions from cnn"*
3	The headline includes the main point of the paragraph, but have a different focus
	e.g., *"health experts condemn donald trump's who funding freeze: 'crime against humanity'"*
4	The headline captures the main point of the paragraph, but contains some information that cannot be inferred from the paragraph
	e.g., *"trump cuts off u.s. funding to who, pending review"*
5	The content of the headline and the paragraph are well aligned
	e.g., *"coronavirus: us to halt funding to who, says trump"*

Figure 5: Sample headlines with explanations on human evaluation aspect scales.

Example A

Article: [No Tiger Woods and no Rory McIlroy in the field is proving no problem in Houston, where record numbers have descended on the small suburb of Humble for the Shell Houston Open.] . . . It's 18 months now since Spieth attracted criticism for giving up on university and joining the PGA Tour as the 809th ranked player in the world. Spieth, only 21, has since risen to number four in the world. Funnily enough, no-one's criticising him anymore. Spectators have flocked to see Texas local boy Jordan Spieth in action at the Shell Houston Open. Spieth's fellow American Phil Mickelson has also **impressed** in **the run up to the Masters** . . . .

Reference: Record numbers have flocked to see local boy Jordan Spieth in action . Spieth and Phil Mickelson have been gearing up for the Masters . Texas-born Spieth quit university in order to concentrate on golf career .

Decoder State Adjustment: [Tiger Woods and Rory McIlroy are not in the field at the Shell Houston Open. Record numbers have descended on the small suburb of Humble.] Phil Mickelson and local boy Jordan Spieth have impressed in the run up to the Masters. Spieth, only 21, has risen to number four in the world. (*reorganize sentences*)

Word Unit Prediction: The Shell Houston Open is the last tournament before the Masters. Phil Mickelson and Jordan Spieth are both **playing well** at **the tournament**. Spieth, only 21, is now number four in the world after 18 months on the PGA Tour. The American has finished in the top seven only twice this season.

Example B

Article: . . . [Raikkonen's contract finishes at the end of the current Formula One season, although there is an option for 2016 providing both parties are in agreement.] The Finn stated this week he has never been happier working with a team in his entire F1 career, although his form to date has not matched that of team-mate Sebastian Vettel. [Kimi Raikkonen has been **urged** to **improve his performances** if he wants to stay at Ferrari.] . . .

Reference: Kimi Raikkonen's deal with Ferrari expires at the end of the season . Team principal Maurizio Arrivabene wants Finn to improve performances . Maranello driver has struggled to replicate team-mate Sebastian Vettel .

Decoder State Adjustment: [Kimi Raikkonen's contract with Ferrari ends at the end of the season. There is an option for him to stay with the Maranello marque in 2016.] Raikk Finnish has said he has never been happier working with a team. [But team principal Maurizio Arrivabene has urged him to improve.]

Word Unit Prediction: Kimi Raikkonen is out of contract with Ferrari at the end of the season. The Finn has said he has never been better at a team in his F1 career. But team-mate Sebastian Vettel has been in better form so far this season. Maurizio Arrivabene has said that **results will have to go up** if he wants to be at Ferrari for another season.

Example C

Article: Three British citizens arrested in the United Arab Emirates after they were found plane spotting near Fujairah airport are to be released Monday, their lawyer, Nasser al-Hashem, tells CNN. The three have been in jail since February 22. . . . No charges were filed, there will be no travel ban, and the men will not face **deportation**, he said. [Conrad Clitheroes, 54, and Gary Cooper, 45,] were on a five-day visit to the UAE from Manchester when they were arrested. The third man, [Neil Munro], is a British national who lives in the UAE. As a hobby, plane spotters view and photograph aircraft around the world.

Reference: Three British men won't be charged or deported, their lawyer says. They were arrested after plane spotting near Fujairah airport and have been in jail since February 22.

Decoder State Adjustment: Three Britons were arrested in the United Arab Emirates in February. They were found plane spotting near Fujairah airport. The men will be released Monday, their lawyer says. No charges were filed, and the men will not face deportation.. The three men were on a five-day visit to the UAE from Manchester. (*omit the name*)

Word Unit Prediction: The three have been in jail since February 22. They were arrested near Fujairah airport in the United Arab Emirates. The three will be released on Monday. No charges will be brought against them, and they will not **be sent back to** the U.K. or other countries.

Figure 6: Examples of document summarization on CNN/DM dataset. We highlight lexical simplifications with blue colors. Texts removed for simplification by decoder state adjustment are in **gray** or explained in *italics*.

Example A		
	Paragraph: Acting chief of staff Mick Mulvaney says President Trump willing to accept a barrier made of steel	
REFERENCE	mulvaney: saturday shutdown meeting 'did not make much progress'	mick mulvaney: trump willing to take concrete wall 'off the table'
RERANKING	mick mulvaney says trump willing to accept a barrier made of steel	mick mulvaney: trump willing to accept steel barrier
LBLCTRL	mick mulvaney: trump willing to accept barrier made of steel	mick mulvaney: trump willing to accept barrier made of steel
IDEOL.SCORER	trump's budget proposal would increase the number of **military contractors** in the us	trump wants to build a wall, **and he's willing to pay for it**
DYNAMIC WORDU	trump wants a border wall, **but it's not all about the wall**	mick mulvaney: trump willing to accept 'steel' border wall
Example B		
	Paragraph: Rep. Paul Ryan accused President Barack Obama of emboldening Iran and those storming U.S. embassies abroad while curtailing individual freedoms at home, during a speech here to a gathering of religious conservatives.	
REFERENCE	paul ryan hits obama on national security: if we project weakness, they come	ryan to values voters: "american foreign policy needs moral clarity"
RERANKING	paul ryan accuses obama of emboldening iran, protesters	paul ryan: obama emboldens iran **healthcare bill**
LBLCTRL	paul ryan: obama emboldening iran	ryan: obama emboldened iran, embassy protesters
IDEOL.SCORER	paul ryan accuses obama of emboldening iran, protesters at **religious conservatives**' gathering	ryan: obama emboldening iran, protesters while **curtailing freedoms at home**
DYNAMIC WORDU	paul ryan to **religious conservatives**: obama has 'emboldened' iran	paul ryan: obama has 'emboldened' iran, protesters
Example C		
	Paragraph: The FBI on Wednesday issued an extraordinary public statement condemning the Republican push to release a classified memo that alleges surveillance abuses at the Department of Justice.	
REFERENCE	opinion: why trump is so eager to release the nunes memo	trump to declassify infamous fisa memo
RERANKING	the fbi just responded to the **gop**'s push to release the memo	fbi condemns **gop** push to release classified memo
LBLCTRL	the fbi just issued a public statement condemning the release of the **republican** memo	fbi condemns **gop** push to release classified memo
IDEOL.SCORER	the fbi just released a statement condemning the release of the **republican** memo	fbi releases statement condemning release of **russia** memo
DYNAMIC WORDU	fbi condemns **gop** push to release classified memo on **russia**	fbi condemns **gop** push to release memo on **surveillance abuses**

Figure 7: Examples of ideology-controlled headline generation. Best viewed in color. The left panel (shaded in blue) shows headlines generated with control toward the left stance. The right panel (red) shows headlines generated with control toward the right. We highlight words that are commonly used with the corresponding stances in **bold**.