Make Lead Bias in Your Favor: Zero-shot Abstractive News Summarization

Chenguang Zhu[1], **Ziyi Yang**[2], **Robert Gmyr**[1], **Michael Zeng**[1], **Xuedong Huang**[1]
Microsoft Cognitive Services Research Group[1]
Stanford University[2]
{chezhu, rogmyr, nzeng, xdh}@microsoft.com, zy99@stanford.edu

Abstract

Lead bias is a common phenomenon in news summarization, where early parts of an article often contain the most salient information. While many algorithms exploit this fact in summary generation, it has a detrimental effect on teaching the model to discriminate and extract important information. We propose that the lead bias can be leveraged in a simple and effective way in our favor to pre-train abstractive news summarization models on large-scale unlabeled corpus: predicting the leading sentences using the rest of an article. We collect a massive news corpus and conduct careful data cleaning and filtering. We then apply the proposed self-supervised pre-training to existing generation models BART and T5. Via extensive experiments on six benchmark datasets, we show that this approach can dramatically improve the quality of summary and achieve state-of-the-art results for zero-shot news summarization without any fine-tuning. For example, in the DUC-2003 dataset, the ROUGE-1 of BART increases 13.7% after the lead-bias pre-training.

1 Introduction

The goal of text summarization is to condense a piece of text into a shorter version that contains the salient information. Due to the prevalence of news articles and the need to provide succinct summaries for readers, a majority of existing datasets for summarization come from the news domain [1, 2, 3]. However, according to journalistic conventions, the most important information in a news report usually appears near the beginning of the article [4, 5]. While it facilitates faster and easier understanding of the news for readers, this lead bias causes undesirable consequences for summarization models. The output from these models is inevitably affected by the positional information of sentences. For instance, [6] discovers that most models' performances drop significantly when a random sentence is inserted in the leading position, or when the sentences in a news article are shuffled.

Additionally, most current summarization models are fully supervised and require time-consuming and labor-intensive annotations to feed their insatiable appetite for labeled data. For example, the CNN/DailyMail dataset [2] contains 313k articles, where the summaries are written by editors. Therefore, some recent work [7] leverage domain transfer to apply summarization models trained on one dataset to another dataset. But this method may be affected by the domain drift problem and still suffers from the lack of labelled data.

The recent promising trend of pre-trained models [8, 9] proves that massive data can be used in a self-supervised fashion to boost the performance of NLP models. But it remains a challenge how to design pre-training goals for text summarization. In this paper, we put forward a novel method to leverage the lead bias of news articles in our favor to conduct large-scale pre-training of summarization models. The idea is to predict the leading sentences of a news article given the rest

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

content. This immediately renders the large quantity of unlabeled news articles corpus available for training news summarization models.

We employ this pre-training idea on a three-year collection of online news articles. We conduct thorough data cleaning and filtering. For example, to maintain a quality assurance bar for using leading sentences as the summary, we compute the ratio of overlapping non-stopping words between the top 3 sentences and the rest of the article. As a higher ratio implies a closer semantic connection, we only keep articles for which this ratio is higher than a threshold determined via statistical analysis. As a result, in the filtered dataset consisting of 21.4M articles, the leading sentences can be a surrogate summary of good quality.

Inspired by the effectiveness of additional pre-training to adapt language models to domains [10], we apply the lead-bias (LB) pre-training on existing generation models including BART [11] and T5 [12]. The resulting models BART-LB and T5-LB are leveraged in a zero-shot fashion, i.e. directly applied to target tasks without accessing *any* information for fine-tuning. Therefore, the same pre-trained model can be used across various news summarization tasks.

We conduct extensive evaluations on six news summarization datasets. Results show that our models significantly improve the summary quality over the original BART and T5, as well as other zero-shot and unsupervised summarization models. For example, BART-LB outperforms BART by 13.7%, 8.3% and 7.7% in ROUGE-1 on DUC-2003, DUC-2004 and CNN/DailyMail. Also, BART-LB outperforms the zero-shot version of PEGASUS [13] by 7.6%, 6.9%, and 1.8% in ROUGE-1 on CNN/DailyMail, XSum and Gigawords, respectively.

2 Related work

2.1 Unsupervised Text Summarization

Traditional abstractive summarization models [14, 15] leverage labeled summaries which require lots of manual work. Therefore, unsupervised abstractive summarization (UAS) models aim to learn to summarize from articles alone. Among these approaches, [16] tries to reconstruct the input article from the generated summary. [17] adopts de-noising autoencoders for sentence compression. [18] leverages reinforcement learning and adversarial training to improve the readability of generated summaries. [19] projects articles and sentences into a common space from which reconstruction can be conducted. [20] employs theme modeling and a de-noising autoencoder to enhance the quality of summaries. Although [20] also leverages lead bias for pre-training, it is trained from scratch and is employed in unsupervised summarization, while our work applies lead bias to pre-trained generation models for domain adaptation and is evaluated in a zero-shot setting.

In general, these models require updating on the articles in the training set of target tasks, and the model trained on one dataset may not be suitable for another dataset. In comparison, our model is strictly zero-shot in that it does not access any information in the training set of downstream tasks. So the same model can be applied to multiple news summarization tasks.

2.2 Pre-trained generation models

In recent years, pre-training language models have proved to be quite helpful in natural language generation tasks [21, 11, 12, 9]. Built upon large-scale corpora, these models employ self-supervised learning such as de-noising autoencoder and masked language model to learn effective representations.

In theory, any generation model can be directly used for zero-shot abstractive summarization (ZAS). Thus, recent large-scale pre-trained generation models have been applied to ZAS. For example, the GPT-2 model [9] can produce summaries given an article appended with *TL;DR:*. PEGASUS [13] uses gap sentences generation (GSG) and masked language model (MLM) as pre-training objectives. Although GSG includes top-m sentences as decoder targets, we apply a statistical cleaning and filtering mechanism to the pre-training data. And out models significantly outperform PEGASUS in multiple datasets.

3 Pre-training with Lead Bias

News articles usually follow the convention of placing the most important information early in the content, forming an inverted pyramid structure [4, 5, 6]. This positional bias brings lots of difficulty for models to extract salient information from the article [6].

We propose that the lead bias in news articles can be leveraged in our favor to train an abstractive summarization model without labeled summaries. Given a news article, we treat the top three sentences, denoted by Lead-3, as the target summary, and use the rest of the article as news content, denoted by Rest. The goal of the summarization model is to produce Lead-3 given the text from Rest.

Thus, we collect three years of online news articles from June 2016 to June 2019. We filter out articles which overlap with any of the evaluation datasets. However, not all leading sentences are suitable for a summary of the article. An indiscriminative usage of all the data may hurt the model's summarization capability. Thus, one should carefully examine and clean the source data to ensure the quality of the leading sentences as summary.

First, to ensure that the summary is concise and the article contains enough salient information, we only keep articles with 10-150 words in the top three sentences and 150-1200 words in the rest, and that contain at least 6 sentences in total.

Second, we remove articles whose top three sentences may not form a relevant summary. For this purpose, we compute the portion of non-stopping words in the top three sentences that are also in the rest of the article. A higher portion implies that the summary is representative and has a higher chance of being inferred by the model using the rest of the article. To verify, we compute the overlapping ratio of non-stopping words between human-edited summary and the article in CNN/DailyMail dataset, which has a median value of 0.841. This median drops moderately to 0.778 between the summary and Rest. And the ratio we use for filtering, i.e. that between Lead-3 and Rest, has a median value of 0.471. Thus, we need to filter out articles with low overlapping ratio between Lead-3 and Rest to use Lead-3 as a surrogate summary. We end up choosing a threshold of 0.65 to strike a balance between the quality of Lead-3 and the size of training data. This filters out 83.2% of the collected data. The retained data for pre-training has a median overlapping ratio of 0.734 between Lead-3 and Rest.

After the filtering, we end up with 21.4M news articles. The average number of words in Lead-3 is 60.0 and the average number of words in Rest is 602.5. Therefore, the data for pre-training contains 14.2 billion words.

We initialize the two versions of our model with BART-Large [11] and T5-Large [12] respectively. Each model is pre-trained for 1 epoch as we found that further pre-training does not bring additional gain. The pre-training takes 47 hours on 32 V-100 GPUs. We denote the pre-trained models as **BART-LB** and **T5-LB**. More details are described in Appendix A.

4 Experiments

4.1 Settings

We evaluate our model on 6 benchmark news summarization datasets: DUC-2003 [22], DUC-2004 [22], XSum [3], CNN/DailyMail dataset [1], the New York Times (NYT) [2] and Gigaword [23]. We use the ROUGE F1 score [24] as the evaluation metric for all datasets except NYT, where ROUGE recall score is used. In NYT, the generated summary is truncated to the same length as the ground-truth summary. In DUC-2003/2004, the generated summary is truncated to 75 characters.

We include both unsupervised summarization models and zero-shot models as baselines. The unsupervised baselines include SEQ^3 [25], Brief [18] and TED [20]. The zero-shot baselines include PEGASUS (zero-shot) [13], GPT-2 [9], BART-Large [11] and T5-Large [12]. To compare with leading sentences, we follow the previous results to use Lead-8 for Gigaword, Lead-1 for XSum, leading 75 characters for DUC-2003/DUC-2004 and Lead-3 for all the other datasets.

Model	**DUC-2003**			**DUC-2004**			**XSum**		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
Lead	**21.30**	**6.38**	**18.82**	20.91	5.52	18.20	16.30	1.60	11.95
Unsupervised									
SEQ^3	20.90	6.08	18.55	**22.13**	6.18	**19.30**	/	/	/
Zero-shot									
PEGASUS	/	/	/	/	/	/	19.27	3.00	12.72
BART$_{\text{LARGE}}$	6.69	1.56	5.94	13.58	2.91	12.10	19.26	3.30	14.67
T5$_{\text{LARGE}}$	10.11	2.43	9.25	13.61	2.91	12.23	19.66	2.91	15.31
BART-LB (ours)	20.43	5.80	17.89	21.88	**6.24**	19.22	**26.18**	**7.60**	**20.92**
T5-LB (ours)	20.05	5.62	17.83	21.22	5.92	18.74	26.06	6.77	20.47
Model	**CNN/DM**			**NYT**			**Gigaword**		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
Lead	40.34	**17.70**	36.57	39.58	20.11	35.78	21.86	7.66	20.45
Unsupervised									
SEQ^3	23.24	7.10	22.15	17.85	3.94	19.53	25.39	8.21	22.68
Brief	28.11	9.97	25.41	/	/	/	21.26	5.60	18.89
TED	38.73	16.84	35.40	/	/	/	**25.58**	**8.94**	**22.83**
Zero-shot									
GPT-2	29.34	8.27	26.58	/	/	/	/	/	/
PEGASUS	32.90	13.28	29.38	/	/	/	23.39	7.59	20.20
BART$_{\text{LARGE}}$	32.83	13.30	29.64	32.18	13.90	28.67	22.07	7.47	20.02
T5$_{\text{LARGE}}$	39.68	17.24	36.28	32.78	14.91	29.91	15.67	4.86	14.38
BART-LB (ours)	**40.52**	17.63	**36.76**	37.41	19.60	33.99	25.14	8.72	22.35
T5-LB (ours)	38.47	16.62	35.23	**40.27**	**20.81**	**36.88**	24.00	8.19	21.62

Table 1: ROUGE results on the test set of all datasets. We use ROUGE recall scores in NYT and F1 scores in all other datasets. The highest score in each dataset is marked in bold.

4.2 Results

As shown in Table 1, BART-LB and T5-LB achieve the best overall result in XSum, CNN/DailyMail and NYT. In other datasets, our models outperform all zero-shot baselines and are comparable with unsupervised models and Lead baseline. For instance, BART-LB outperforms the zero-shot version of PEGASUS [13] by 7.6%, 6.9% and 1.8% in ROUGE-1 on CNN/DailyMail, XSum and Gigawords.

The proposed self-supervised pre-training based on lead bias is very effective in improving performance of the underlying pre-trained model. For example, BART-LB improves BART by 13.7%, 8.3% and 7.7% in ROUGE-1 on DUC-2003, DUC-2004 and CNN/DailyMail, respectively. T5-LB improves T5 by 9.9%, 8.3% and 7.6% in ROUGE-1 on DUC-2003, Gigaword and DUC-2004.

Thirdly, BART-LB and T5-LB significantly outperform all unsupervised baselines in CNN/DM and NYT, and achieve very close results on the rest datasets. We argue that although unsupervised models are fine-tuned with articles from target tasks, our models pre-trained on massive news data can achieve similar or better results. Moreover, compared with unsupervised models, a zero-shot model can be directly applied to many news summarization domains, making deployment much more convenient.

Furthermore, BART-LB and T5-LB do not simply learn to copy leading sentences. They outperform the Lead baseline in 5 out of 6 datasets, ranging from 0.18% (CNN/DM) to 9.9% (XSum) more ROUGE-1 points. We show more insights from the results in Appendix B.

5 Conclusions

In this paper, we propose a simple and effective pre-training method for abstractive news summarization. By employing the leading sentences from a news article as its target summary, we turn the problematic lead bias for news summarization in our favor. We then collect a large-scale news corpus and conduct filtering based on statistical analysis. We initialize our model with BART and T5, then further pre-train it using the lead bias. The resulting models outperform all zero-shot baselines and achieve comparable results with unsupervised methods in six benchmark datasets.

References

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

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

[3] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

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

[5] Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. *EMNLP*, 2019.

[6] Matt Grenander, Yue Dong, Jackie C.K. Cheung, and Annie Louis. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. *EMNLP*, 2019.

[7] Ilya Gusev. Importance of copying mechanism for news headline generation. *arXiv preprint arXiv:1904.11475*, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[10] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

[11] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[13] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*, 2019.

[14] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[15] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

[16] Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. Seqˆ 3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. *arXiv preprint arXiv:1904.03651*, 2019.

[17] Thibault Févry and Jason Phang. Unsupervised Sentence Compression using Denoising Auto-Encoders. *arXiv e-prints*, page arXiv:1809.02669, Sep 2018.

[18] Yau-Shian Wang and Hung-Yi Lee. Learning to encode text as human-readable summaries using generative adversarial networks. *arXiv preprint arXiv:1810.02851*, 2018.

[19] Peter J Liu, Yu-An Chung, and Jie Ren. Summae: Zero-shot abstractive text summarization using length-agnostic auto-encoders. *arXiv preprint arXiv:1910.00998*, 2019.

[20] Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. Ted: A pretrained unsupervised summarization model with theme modeling and denoising, 2020.

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

[22] Paul Over, Hoa Dang, and Donna Harman. Duc in context. *Information Processing & Management*, 43(6):1506–1520, 2007.

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

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out,*, 2004.

[25] Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. Seqˆ 3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. *arXiv preprint arXiv:1904.03651*, 2019.

[26] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

A Implementation Details

The batch size is 1,024. We use RAdam [26] as the optimizer, with a learning rate of 3×10^{-4}. The hyper-parameters of our pre-training follow the version of BART and T5 implemented by Huggingface. For example, for both BART-LB and T5-LB, the dropout rate is 0.1 and each input token is represented by a 1024-dim vector. BART-LB has 12 transformer layers and 16 attention heads in both the encoder and decoder. T5-LB has 24 transformer layers and 16 attention heads in both the encoder and decoder.

In downstream summarization datasets, we employ the commonly used hyper-parameters by previous models (e.g. set in the configuration files of BART and T5) on the corresponding tasks. These hyper-parameters are for decoding based on beam search. Table 2 shows the minimum summary length, maximum summary length and beam width for each task.

Dataset	sum. minlen	sum. maxlen	beam width
DUC-2003	6	26	1
DUC-2004	6	26	1
XSum	11	62	6
CNN/DailyMail	56	142	4
NYT	56	142	4
Gigaword	4	24	4

Table 2: Hyper-parameters used on each dataset, including minimum/maximum length of produced summary and beam width. These parameters are determined by the implementation of summarization models from previous literature.

B Insights

Does our model simply copy leading sentences? Since extracting leading sentences from articles as the summary can achieve high ROUGE scores in a number of news summarization datasets, we test whether our pre-trained model simply learns to copy the leading sentences.

Following [14], we compute the ratio of novel n-grams appearing in a model's summary that are not in the leading sentences. A higher ratio indicates a system that is more inclined not to copy the leading sentences from the article.

Figure 1 displays the percentage of novel n-grams not in the article's LEAD-1 sentence in summaries from BART, BART-LB and the reference summary in XSum's test set. Firstly, the lead-bias pre-training increases this ratio by comparing the results from BART and BART-LB. The reason is that during pre-training the model needs to predict LEAD-3 using *the rest* of the article. Simply copying the first few sentences from the rest of the article typically match LEAD-3. Thus, our proposed pre-training enforces the model to learn to comprehend and extract salient information from the whole article. Secondly, the reference summary has the highest ratio of novel n-grams, indicating that

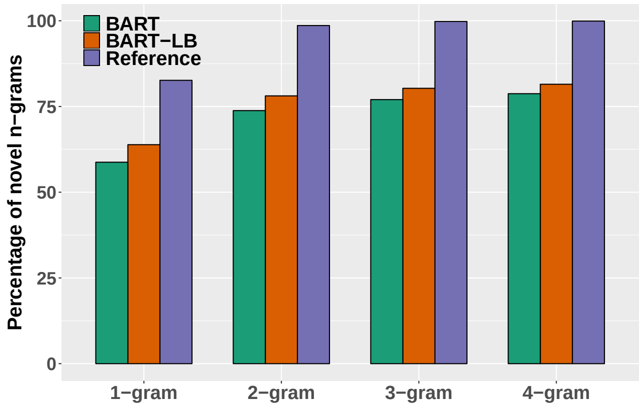

Figure 1: Ratio of novel n-grams, i.e. not in the article's leading sentence, in summaries from BART, BART-LB and the reference summary in XSum's test set.

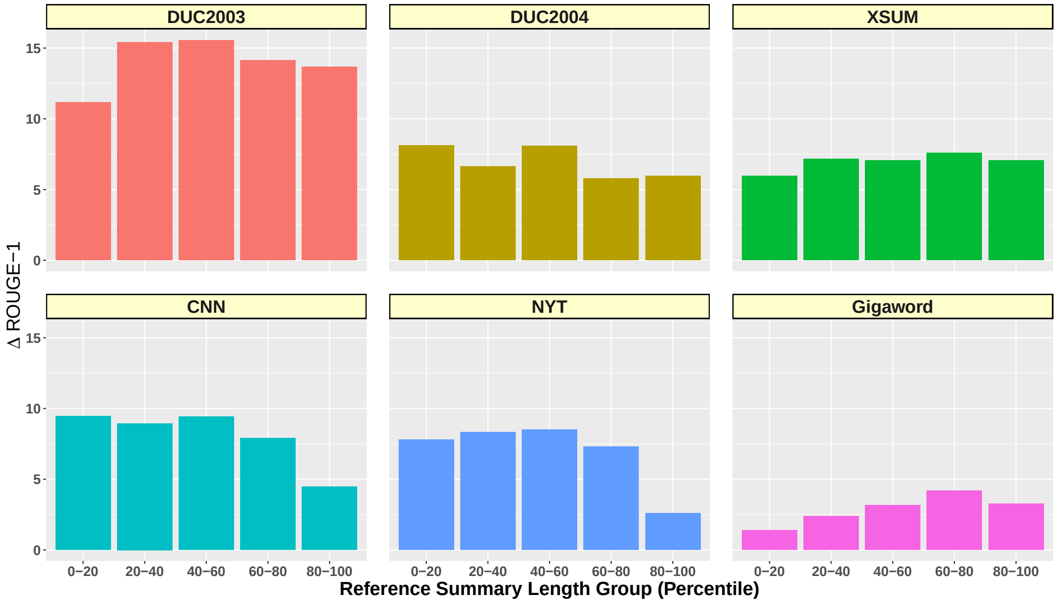

Figure 2: Averaged ROUGE-1 difference between the summaries from BART and BART-LB, grouped by the length of ground-truth summary. For example, 0-20 means the 20% articles with the shortest reference summary in the corresponding dataset.

humans usually summarize the article by reorganizing the language and placing collected important information together, instead of copying original information verbatim.

Effects of summary length. We investigate whether the improvement brought by our lead-bias pre-training is affected by the length of summaries. Thus, we take the $\text{BART}_{\text{LARGE}}$ and BART-LB models, and compute the difference between their ROUGE-1 scores when the reference summary length falls into different percentiles of the dataset: 0-20%, 20-40%, 40-60%, 60-80% and 80-100%.

As shown by Figure 2, in general, the gain of BART-LB over $\text{BART}_{\text{LARGE}}$ is the largest when the reference summary is at 60-80% and 40-60% percentile. Although longer summaries are typically more difficult to generate, the results indicate that our pre-training scheme can help more with producing medium and medium-long summaries.

We attribute the dip at 80-100% to the capacity limit of the summarization model. It becomes harder to achieve a high ROUGE score with a very long ground-truth summary. This problem can be alleviated by developing more powerful generation models to produce long and consistent text.