Improving Factuality of Abstractive Summarization via Contrastive Reward Learning

I-Chun Chern[1] **Zhiruo Wang**[1] **Sanjan Das**[1] **Bhavuk Sharma**[1]
Pengfei Liu[2] **Graham Neubig**[1]
[1] Carnegie Mellon University
[2] Shanghai Jiao Tong University
{ichern, zhiruow, sanjand, bhavuks, gneubig}@cs.cmu.edu
stefanpengfei@gmail.com

Abstract

Modern abstractive summarization models often generate summaries that contain hallucinated or contradictory information. In this paper, we propose a simple but effective contrastive learning framework that incorporates recent developments in reward learning and factuality metrics. Empirical studies demonstrate that the proposed framework enables summarization models to learn from feedback of factuality metrics using contrastive reward learning, leading to more factual summaries by human evaluations. This suggests that further advances in learning and evaluation algorithms can feed directly into providing more factual summaries. Code and human evaluation results will be publicly available at https://github.com/EthanC111/factuality_summarization.

1 Introduction

One major challenge in current abstractive summarization models is how to generate more factual summaries that are consistent with the source text (Li et al., 2022). Various approaches have been proposed to address this challenge, including augmenting the model input (Dou et al., 2021), performing post-processing (Dong et al., 2020; Cao et al., 2020), and modifying the learning algorithms (Cao and Wang, 2021; Liu et al., 2021). In particular, learning-based methods possess the advantage of not requiring modification to the existing model architecture or the addition of new modules.

In the meantime, with the growing interest in aligning learning objectives with evaluation criteria of interest, utilizing feedback of automatic evaluation metrics (Liu et al., 2022) or human preferences (Stiennon et al., 2020) as rewards for fine-tuning abstractive summarization models has gained substantial attention. These methods learn to optimize rewards using techniques such as reinforcement-learning (RL) (Stiennon et al., 2020), minimum risk training (MRT) (Shen et al., 2016; Wieting et al., 2019), and contrastive reward learning (CRL) (Liu and Liu, 2021; Liu et al., 2022).

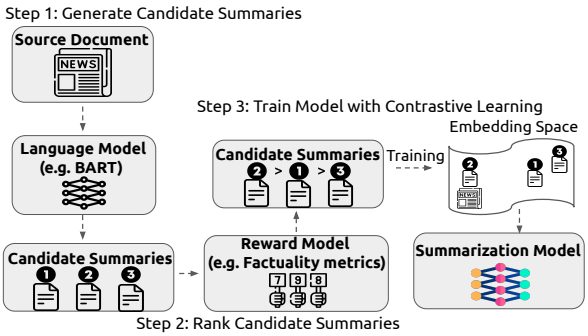

Figure 1: An illustration of our learning framework.

Given the benefits of learning-based methods in improving factuality of abstractive summarization, and recent advancements in factuality metrics for detecting factual inconsistencies in generated summaries, it is of interest to apply reward learning to enforce models to learn from feedback of factuality metrics to improve the factuality of abstractive summarization models. We aim to investigate the following questions in this paper - **Q1:** Can contrastive reward learning effectively utilize existing factuality metrics to improve the factuality of abstractive summarization models? **Q2:** Can the improvement in factuality be reflected in human evaluation studies?

In this paper, we propose a contrastive reward learning framework that enables abstractive summarization models to directly learn from feedback of factuality metrics in a sample-efficient manner. In contrast to other contrastive learning frameworks (Cao and Wang, 2021; Liu et al., 2021), our proposed framework does not rely on the complex construction of negative samples. Instead, similar to (Liu et al., 2022), all candidate summaries used for contrastive learning are generated from pretrained sequence-to-sequence models (Lewis et al., 2020; Zhang et al., 2020) using diverse beam search (Vi-

Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), pages 55–60

jayakumar et al., 2018). Additionally, our framework also incorporates the use of quality metrics to provide more fine-grained information on the ranking (positive / negative) of candidate summaries. Specifically, we investigate learning from the rewards of two factuality metrics: BARTScore (Yuan et al., 2021) and DAE (Goyal and Durrett, 2021). Through automatic and human evaluation studies, we demonstrate that our framework enables summarization models to generate significantly more factual summaries.

2 Contrastive Learning from Factuality Rewards

2.1 Contrastive Learning for Abstractive Summarization

Abstractive Summarization Given a source document D, the summarization model learns a generative model g_θ, that converts the source document D into a summary S:

$$S = g_\theta(D) \tag{1}$$

MLE Loss Given a training sample pair $\{D, S^r\}$ consists of source document D and reference summary S^r (note that S^r consists of L tokens, $S^r = \{s_1^r, \cdots, s_j^r, \cdots, s_L^r\}$), the MLE loss $\mathcal{L}_{\text{mle}}$ aims to maximize the likelihood of reference summary S^r given the source document D:

$$\mathcal{L}_{\text{mle}} = \log p_{g_\theta}(S^r|D) = \sum_{j=1}^{L} \log p_{g_\theta}(s_j^r|D, s_{<j}^r) \tag{2}$$

where $s_{<j}^r = \{s_0^r, \cdots, s_{j-1}^r\}$ and s_0^r is a predefined start token.

Despite its effectiveness in enforcing generated summaries to align with the reference summaries, the MLE loss is not aware of the *quality* (evaluated by some quality metric M) of the generated summaries. To address this issue, we introduce a contrastive loss (Liu et al., 2022).

Contrastive Loss Given a training sample pair $\{D, S^r\}$, and that S_i, S_j are candidate summaries generated from a pre-trained model given D, and that $M(S_i) > M(S_j)\ \forall i, j, i < j$ [1], the contrastive loss is defined as:

$$\mathcal{L}_{ctr} = \sum_i \sum_{j>i} \max(0, f(S_j) - f(S_i) + \lambda_{ij}) \tag{3}$$

Note that $\lambda_{ij} = (j - i) \times \lambda$ is the rank difference between two candidates times a constant λ (usually set as 1) [2] and that $f(S)$ is the length-normalized estimated log-probability given by:

$$f(S) = \frac{\sum_{t=1}^{l} \log p_{g_\theta}(s_t|D, S_{<t})}{|S|^\alpha} \tag{4}$$

where α is a constant.

Intuitively, the contrastive loss penalizes any discoordination between the length-normalized estimated log-probability and the quality metric evaluation (i.e., when $f(S_j) > f(S_i)$ but $M(S_i) > M(S_j)$). The quality metric M could be any evaluation criteria, including automatic evaluation metrics (Lin, 2004; Yuan et al., 2021; Goyal and Durrett, 2021), or human preferences (Ouyang et al., 2022).

Combined Loss The combined loss used for fine-tuning is described by Equation 5.

$$\mathcal{L}_{com} = \mathcal{L}_{\text{mle}} + \gamma \mathcal{L}_{ctr} \tag{5}$$

where $\mathcal{L}_{\text{mle}}$ is the MLE loss given in Equation 2, $\mathcal{L}_{ctr}$ is the contrastive loss given in Equation 3, and γ is the weight of contrastive loss. Summarization models fine-tuned with $\mathcal{L}_{com}$ is referred as CRL-COM.

2.2 Reward from Factuality Metrics

We use two factuality metrics as quality metrics M for use in the contrastive loss described in Equation 3.

BARTScore (Yuan et al., 2021)'s factuality score is calculated as the log-likelihood of the summary given the source calculated from a reference-free version of BARTScore.

DAE (Goyal and Durrett, 2021) is calculated as the softmax output of the least-factual dependency-arc inside the sentences in the summary.

These two metrics were chosen for relative computational efficiency, as they are evaluated many times in the training process. [3]

[1] M could be reference-free (e.g., BARTScore, DAE) or reference-based (e.g., ROUGE) metric. If M is a reference-free metric, then $M(S_i) = M(S_i, D)$; if M is a reference-based metric, then $M(S_i) = M(S_i, S^r)$

[2] The magnitude of contrastive loss can be directly regulated through the weight of contrastive loss γ, so we simply set λ equal to 1.

[3] In contrast, QA-based factuality metrics are computationally inefficient (Laban et al., 2022). As a result, they are less feasible for use in reward-learning settings.

3 Experiments

3.1 Experimental Setup

Driven by the two research questions presented in the introduction, we train two factuality-driven summarization models, namely CRL-COM (B) and CRL-COM (D), trained from contrastive reward learning using BARTScore and DAE as quality metrics, respectively. A baseline summarization model CRL-COM (R) is also trained from contrastive reward learning using ROUGE as quality metric. Note that commonly used n-gram based metrics, including ROUGE (Lin, 2004), have been shown to have a low correlation with human evaluations, particularly on factuality perspective (Falke et al., 2019; Durmus et al., 2020). Thus, we focus on evaluating the factuality of CRL-COM (B) and CRL-COM (D) compared to CRL-COM (R), with the hypothesis that CRL-COM (B) and CRL-COM (D) should be capable of generating more factual summaries compare to CRL-COM (R).

Datasets: We use two abstractive summarization datasets – CNN/Daily Mail (CNNDM) dataset (Hermann et al., 2015; Nallapati et al., 2016) and the XSUM dataset (Narayan et al., 2018). CNNDM summaries tend to be more extractive and are composed of multi-sentence summaries, while XSUM summaries are more abstractive and are composed of single-sentence summaries.

Models: Following the setting outlined in (Liu et al., 2022), we fine-tuned a pre-trained BART model (Lewis et al., 2020) on the CNNDM dataset and a pre-trained PEGASUS (Zhang et al., 2020) model on the XSUM dataset.

Implementation and Fine-tuning Details: The combined loss (with weight of the contrastive loss $\gamma = 100$) described in Equation 5 is used to fine-tune the pre-trained models. Following (Liu et al., 2022) few-shot fine-tuning learning paradigm, we sampled 1000 training samples from each dataset for few-shot fine-tuning. A constant learning rate of 10^{-5} and 10^{-4} was applied to the fine-tuning process for the CNNDM and XSUM datasets, respectively, in order to facilitate fast convergence. For each dataset, we fine-tuned three models using three different quality metrics: ROUGE (R), BARTScore (B), and DAE (D), designated as CRL-COM (R), CRL-COM (B), and CRL-COM (D), respectively. During validation, we employed the same quality metric used for fine-tuning for early stopping.

Automatic Evaluation Each model is evaluated on three metrics: ROUGE (with variants ROUGE-1, ROUGE-2, ROUGE-L), BARTScore, and DAE.

Human Evaluation To objectively evaluate the factual consistencies of the generated summaries from each model, we randomly sampled 100 samples from CNNDM and 200 samples from XSUM for human evaluation. We assess each summary from three different perspectives: Factuality (FAC), Coherence (COH), and Relevance (REL), with a particular emphasis on factuality. The assessment follow similar guidelines as in (Liang et al., 2022; Fabbri et al., 2021). The evaluation guidelines provided to the annotators are listed in Table 1. An expert annotator is involved in the human evaluation studies.

3.2 Results and Analysis

Contrastive reward learning can enforce models to learn from feedback of factuality metrics Driven by **Q1**, we observe that results from automatic evaluation presented in Table 2 indicate that contrastive reward learning enables abstractive summarization models to develop in a direction that aligns with existing factuality metrics.

Learning from factuality metrics improves factuality of abstractive summarization. Driven by **Q2**, we observe that results from human evaluation presented in Table 2 indicate that on both datasets, CRL-COM (B) and CRL-COM (D) exhibit superior performance in terms of factuality compared to CRL-COM (R). This suggests that while learning from factuality metrics such as BARTScore and DAE may potentially result in sacrificing the performance of the models on ROUGE scores, the resulting models can generate more factually consistent summaries. In other words, summaries with higher BARTScore or DAE scores but lower ROUGE scores tend to be more factually consistent with the source article compared to those with lower BARTScore or DAE scores but higher ROUGE scores. This further supports the assertion that BARTScore and DAE are effective at capturing factual information.

Learning from factuality metrics did not sacrifice coherence and relevance. According to human evaluations, the summaries generated by CRL-COM (B) and CRL-COM (D) showed comparable coherence and relevance to those generated

Perspective	Guidelines
Factuality (FAC)	If all the information and claims inside the summary are included in the source article, assign a binary score of 1 ; otherwise, assign a binary score of 0.
Coherence (COH)	On a Likert scale of 1 (worst) to 5 (best), assign a score based on how well the relevant information is coordinated and organized into a well-structured summary.
Relevance (REL)	On a Likert scale of 1 (worst) to 5 (best), assign a score based on the extent to which the summary includes only important information from the source article.

Table 1: Guidelines for human evaluation studies

System	Automatic Evaluation					Human Evaluation		
	R-1	R-2	R-L	B	D	FAC	COH	REL
CNNDM								
CRL-COM (R)	**45.75**	**21.87**	**42.27**	-1.43	36.28	0.76	4.00	**4.17**
CRL-COM (B)	41.07	18.15	36.63	**-0.78**	88.92	**0.99**	**4.05**	3.96
CRL-COM (D)	42.20	19.21	38.19	-0.80	**89.48**	**0.99**	4.03	4.04
XSUM								
CRL-COM (R)	**47.28**	**24.14**	**38.78**	-2.42	32.75	0.38	3.52	3.25
CRL-COM (B)	41.85	19.38	33.46	**-1.87**	37.48	**0.51**	**3.73**	**3.50**
CRL-COM (D)	44.38	22.16	36.57	-2.38	**40.91**	0.50	3.62	3.29

Table 2: Results of each system on CNNDM and XSUM dataset. Note that R stands for ROUGE, B stands for BARTScore, and D stands for DAE.

by CRL-COM (R). This suggests that BARTScore and DAE has comparable abilities to ROUGE in terms of measuring coherence and relevance.

4 Related Work

4.1 Factuality Metrics for Abstractive Summarization

Various factuality metrics assess the factual consistency between a summary and its corresponding source document. QA-based factuality metrics leverage question generation (QG) models to generate questions from the summary and question answering (QA) models to answer those questions, given both the source and summary (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021; Fabbri et al., 2022). Factuality is then evaluated based on the alignment between the answers from the source and summary. Another class of metrics, entailment-based factuality metrics (Kryscinski et al., 2020; Goyal and Durrett, 2021; Laban et al., 2022), evaluates whether all the information in the summary is entailed by the source document. Recent studies on leveraging pre-trained language model as evaluation (Yuan et al., 2021) also achieve competitive performance on evaluating factuality.

4.2 Improving Factuality of Abstractive Summarization via Contrastive Learning

Several contrastive learning frameworks have been proposed to enable models to learn factuality from positive samples (such as reference summaries) and negative samples (such as edited reference summaries and system generated summaries). For example, CLIFF (Cao and Wang, 2021) and CO2Sum (Liu et al., 2021). Both of which are similar in nature but CO2Sum employs more sophisticated methods for negative sample construction.

5 Conclusion

In this work, we present a simple contrastive reward learning framework that enforces abstractive summarization models to learn from feedback of existing factuality metrics. Empirical studies demonstrate the effectiveness of this approach, showing that abstractive summarization models that learn from factuality metric feedback through contrastive reward learning can generate more factual summaries without sacrificing coherence or relevance. This suggests that further advancements in the reward learning paradigm and factuality metrics can facilitate the development of more factually consistent abstractive summarization models.

6 Limitations

While we have included two distinctive dataset (CNNDM and XSUM) in our experiments, more non-news datasets could be included in future studies. Other possibilities for future work include comparing the capability of RL-based reward learning and contrastive reward learning in improving the factuality of abstractive summarization models.

7 Ethics Statement

Even though some of the investigated systems may achieve a high level of factuality on the CNNDM dataset, this does not guarantee that they can be used as off-the-shelf factual consistent summarization models. Thorough evaluation should be conducted before using these models in high-stakes settings to ensure their reliability.

Acknowledgements

We would like to thank Yixin Liu for helpful discussion on BRIO. We would also like to thank Tanya Goyal for helpful discussion on DAE.

References

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

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

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

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

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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

Wei Liu, Huanqin Wu, Wenjing Mu, Zhen Li, Tao Chen, and Dan Nie. 2021. Co2sum: Contrastive learning for factual-consistent abstractive summarization. *arXiv preprint arXiv:2112.01147*.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

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

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

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