Improving Factuality of Abstractive Summarization without Sacrificing Summary Quality

Tanay Dixit * **Fei Wang** **Muhao Chen**
Indian Institute of Technology Madras University of Southern California
dixittanay@gmail.com {fwang598,muhaoche}@usc.edu

Abstract

Improving factual consistency of abstractive summarization has been a widely studied topic. However, most of the prior works on training factuality-aware models have ignored the negative effect it has on summary quality. We propose EFACTSUM (i.e., **E**ffective **Fact**ual **Sum**marization), a candidate summary generation and ranking technique to improve summary factuality without sacrificing summary quality. We show that using a contrastive learning framework with our refined candidate summaries leads to significant gains on both factuality and similarity-based metrics. Specifically, we propose a ranking strategy in which we effectively combine two metrics, thereby preventing any conflict during training. Models trained using our approach show up to 6 points of absolute improvement over the base model with respect to FactCC on XSUM and 11 points on CNN/DM, without negatively affecting either similarity-based metrics or absractiveness.[1]

1 Introduction

Although recent methods have made significant improvements in abstractive summarization (Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020), they do still lack a very critical component - factual consistency. Recent works (Cao et al., 2020; Kryscinski et al., 2019; Maynez et al., 2020) have shown that a majority of the model-generated summaries are unfaithful and suffer from a wide range of hallucination (Tang et al., 2022). Making summarization models factually consistent is critical for its trustworthiness in real-world applications.

Recent studies have made several attempts to improve factuality of abstractive summarization by either modifying the maximum likelihood estimation (MLE) training objective (Cao and Wang, 2021; Goyal and Durrett, 2021), directly optimizing factuality metrics using reinforcement learning (Cao et al., 2022) or improving the quality of the training data (Goyal and Durrett, 2021; Nan et al., 2021a). However, most of these works have reported a negative relationship between factual consistency and summary quality[2]. For example, Goyal and Durrett (2021) improve factuality at a cost of a 6-point drop in ROUGE-L, Wan and Bansal (2022) also observe a 2-point drop in ROUGE-L. Prior approaches have also optimized factuality at the cost of abstractiveness (Ladhak et al., 2022). This leads to a critical question: *Can we improve the factuality of summarization without the cost on the summary quality?*

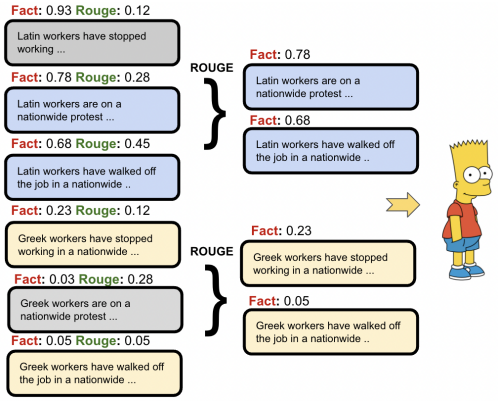

Figure 1: Overview of our approach. For a given article, we generate a number of summaries that can be either factual (blue) or non-factual (yellow). Grey summaries are filtered out. We select a balanced set using ROUGE and then finally train the model to rank them based on the factuality score.

To this end, we propose EFACTSUM (i.e. **E**ffective **Fact**ual **Sum**marization): A candidate summary generation and ranking technique for contrastive summarization training (Fig. 1) that not only achieves significant gains in factuality of abstractive summarization but also improves the sum-

* This work was done when the first author was visiting the University of Southern California.

[1] Code is available at https://github.com/tanay2001/EFactSum.

[2] summary quality as measured by metrics like ROUGE, BERTScore, etc.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 902–913

mary quality. Unlike prior works which often sacrifice summary quality for improving faithfulness, we take an alternative approach to improve both faithfulness and summary quality. We make use of the fine-tuning strategy by Liu et al. (2022) and make key modifications to the ranking process. As depicted in Fig. 1 we start with generating a number of candidate summaries using existing fine-tuned models. Using these summaries, we select a subset by effectively combining two evaluation metrics of the two different criteria (§2), thus avoiding optimizing one at the cost of the other. This technique helps obtain gains over methods that simply optimize one metric (§3.4). The promising results by EFACTSUM on XSUM and CNN/DM have shown consistent improvements in both aspects over strong baselines, demonstrating effectively enhanced summarization factuality without sacrificing the quality.

2 Approach

Given a document (D), the task of summarization seeks to generate its summary (S) that satisfies some conditions like factuality, coherence, etc. The standard fine-tuning process involved the use of Maximum Likelihood Estimation (MLE). Inspired by Liu et al. (2022), in addition to the cross-entropy loss, we incorporate a contrastive loss that encourages models to provide a higher probability mass to the more factual summaries. Formally, for every training document D and a ranked list of the most probable candidate summaries $[S_1, S_2, \ldots S_n]$, the model learns to rank the summaries according to the factuality score. To achieve this, we make use of the following loss:

$$\mathcal{L}_{CL} = \sum_{i} \sum_{j>i} \max(0, f(S_j) - f(S_i) + \lambda_{ij}), \quad (1)$$

where S_i and S_j are two different candidate summaries and S_i ranks higher than S_j, $\lambda_{ij} = (j-i)*\lambda$ is a rank-based margin, and $f(.)$ is the estimated log-probability normalized by length:

$$f(S) = \frac{\sum_{t=1}^{l} \log p_{g_\theta}(s_t | D, S_{<t}; \theta)}{|S|^\alpha}. \quad (2)$$

Candidate Set Generation. To generate the candidate summarization set $\{S_i\}$, we make use of an existing model and sample summaries using beam search (Vijayakumar et al., 2018). We observe that just using the model trained with cross-entropy leads to generating a number of unfaithful summaries. In order to generate more faithful summaries, we make use of factually improved models.

Ranking Strategy. Since our primary goal is to optimize factuality without adversarially affecting summary quality, we need to consider two metrics while deciding the ideal ranking. In order to measure the factuality of S_i, we choose FactCC (Kryscinski et al., 2020) because it correlates well with human judgments of faithfulness (Pagnoni et al., 2021) and it is also computationally more efficient than other question-answering based metrics (Scialom et al., 2021). To measure the summary quality, we use the popular ROUGE metric (Lin, 2004). Now, amongst the set of candidate summaries that have been scored to be faithful, we further choose the top m summaries that have the highest ROUGE score. We select the set of unfaithful summaries in the same way just that we choose the m summaries with the lowest ROUGE scores. This technique of incorporating two evaluation metrics helps overcome the inherent conflict (Chaudhury et al., 2022). We highlight the importance of the proposed steps in §3.4. At last, these $2m$ summaries are used in creating the ranked list of candidate summaries for each article in the training set. The intuition behind this approach is that since the FactCC scores are not confidence scores, summaries from only one set can not provide sufficient supervision signals. Instead, training the model with balanced summaries from both sets would be beneficial.

Finally, our training objective combines the cross-entropy loss and our contrastive loss

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \gamma \mathcal{L}_{CL}, \quad (3)$$

where γ is the weight of the contrastive loss.

3 Experiments

We state the experimental setup in §3.1 and report the results in §3.2, followed by an abstractiveness analysis in §3.3. In §3.4, we analyze the importance of the various components in our approach.

3.1 Experimental Settings

Datasets. To understand the effectiveness of EFACTSUM, we make use of two widely-used news summarization datasets, XSUM (Narayan et al., 2018) and CNN/DM (Hermann et al., 2015).

Model	Summ. Quality R-1	R-L	BS.	Factuality FactCC	DAE ↓
XSUM					
PEGASUS	47.07	39.26	89.19	24.33	0.426
BRIO	48.69	40.13	90.87	21.47	0.452
FASum	29.72	23.29	88.57	26.08	0.616
DAE	38.63	30.22	88.44	26.66	0.462
CLIFF	46.33	38.27	88.96	24.54	**0.386**
EFACTSUM	**47.24**	**39.45**	**89.79**	**30.48**	0.417
CNN/DM					
BART	43.04	39.41	87.21	49.07	0.049
BRIO	47.53	44.02	89.12	30.35	0.093
FASum	40.40	36.97	88.23	51.17	0.046
CLIFF	44.14	40.72	**88.82**	51.84	0.047
EFACTSUM	**44.37**	**40.92**	88.36	**60.74**	**0.041**

Table 1: Results of models fine-tuned on the XSUM and CNN/DM. R-1: Rouge-1 , R-L: Rouge-L , BS: BERTScore. For DAE smaller the better the score. Models perform significantly better than the PEGASUS/BART model ($p < 0.05$). The best result for factuality-aware training methods is **bolded**. Overall best score per metric is underlined.

Baselines. In addition to models fine-tuned using *cross-entropy* and competitive fine-tuning techniques: **BRIO** (Liu et al., 2022), we compare EFACTSUM with prior works that have modified the fine-tuning process to improve factuality, including (1) **CLIFF** (Cao and Wang, 2021) which uses contrastive learning to train summarization models to differentiate between consistent and hallucinated summaries, (2) **FASum** (Zhu et al., 2021) that modifies the Transformer architecture by incorporating knowledge graphs for factual consistency, and (3) **DAE** (Goyal and Durrett, 2021) that masks out the nonfactual tokens during training. This comparison is only available for the XSUM dataset.

Metrics. To evaluate factuality, we make use of FactCC (Kryscinski et al., 2020), a popular metric that uses a BERT-based metric to measure whether the generated output is faithful. We also consider DAE (Goyal and Durrett, 2020), a textual-entailment-based metric that correlates well with human judgment of factuality (Tang et al., 2022). It uses an arc entailment model to evaluate the factuality of a summary. We make use of the token-level score in order to complement the sentence-level scores from FactCC. For quality assessment, we use ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) to evaluate the summary against the reference.

Implementation Details. We use CLIFF and cross-entropy trained models to generate the candidate set of summaries $(S_1, S_2, ..., S_n)$. We use $n = 6$ and only retain those training articles that contain at least 2 factual and non-factual candidate summaries. Using this new subset of training data, we fine-tune BART-Large (Lewis et al., 2020) on CNN/DM and PEGASUS (Zhang et al., 2020) on XSUM. More details are in Appx. §A.

3.2 Main Results

We report the results of the model fine-tuned using our approach in Tab. 1. Outputs of models fine-tuned using our strategy are presented in Tab. 2 and Appx. §C. Overall we can observe the proposed EFACTSUM leads to improvements on both the factuality metrics while preserving or improving the performance on reference-based similarity metrics.

For XSUM, EFACTSUM achieves a notable relative gain of 25% on FactCC and 3% on DAE (token) in comparison to PEGASUS while simultaneously showing non-trivial gains on both ROUGE and BERTScore. Although EFACTSUM is trained to optimize FactCC, it also does well on the other evaluation metric, thus pointing out that the training process does not exploit any biases related to the evaluation metrics. One should note that although CLIFF does better on DAE, it is sacrificing summary quality. A similar story holds for CNN/DM also where EFACTSUM achieves a relative gain of 20% and 16% on FactCC and DAE respectively. Unlike some of the prior works, this gain in factuality has not come at a cost of summary quality or abstractivness (§3.3). Although BRIO outperforms our approach on ROUGE and BERTScore, it substantially decreases factuality score, which is not desirable. Our approach aims to strike a balance between factuality and summary quality.

3.3 Factuality vs Abstractiveness Tradeoff

Ladhak et al. (2022) show that it is naively possible to increase the factuality of generated summaries by increasing extractiveness (decreasing abstractiveness). Hence we analyze the extractiveness level of the generated summaries to understand if our method suffers from this tradeoff. Along with the extractiveness scores (Grusky et al., 2018), we compute the MINT (Metric for lexical INdependence of generated Text) scores and the abstractiveness-adjusted metrics scores (Dreyer et al., 2023). Fig. 2 depicts the extractiveness levels for the various summarization systems. Scores are

System	Summary	Article
Base.	The number of migrants and refugees arriving on the Greek island of Lesbos has halved in the past week.	Lesbos used to get more than 5,000 a day. On Monday there were just four. But with Europe's borders closed, more than 50,000 migrants remain in Greece waiting for a decision about their futures. . . . But here she is in Moria, once a transit camp for migrants, now **since the EU deal with Turkey**, a detention centre, run by central government. . . . It is another sign of how Greece was simply overwhelmed by the numbers who came, while itself in the middle of an economic crisis. Most of those who arrived before March 20, the start of the EU-Turkey agreement, are free to come and go, but cannot leave the island. Those who came after that date are locked in, waiting for a decision . . .
Ours	The number of migrants arriving on the Greek island of Lesbos has halved since the EU struck a deal with Turkey to stem the flow.	
Base	Goldman Sachs will no longer conduct face-to-face interviews with students applying for analyst jobs.	The US investment bank will switch to video interviews with first-round undergraduate candidates from next month. . . . Goldman hoped the move will allow it to find students who do not attend top-tier US universities. . . . **It will still conduct second-round interviews in person**. The shift will not affect business schools or professional hires, but is part of a broader move by Goldman to use technology in the hiring process. The new method will include structured interviews, which the bank said will allow for greater comparisons between candidates . . .
Ours	Goldman Sachs is changing the way it hires students.	
Base	The pilot of a Turkish military jet has died after it crashed in the south-west of the country, state media report.	The plane was flying over the Amanos Mountains in the southern province of Osmaniye on Monday when it lost radio contact, Anatolia news agency said. . . . Rescuers found the pilot's body near to the wreckage of the aircraft. Osmaniye Governor Celalettin Cerrah had earlier announced that a cockpit window and some other pieces of the aircraft had been found in the Caksir area. . . People living around the village of Yarpuz, about 25km (16 miles) north of the **Syrian border, said that they had heard a loud bang like an explosion**, according to local media A Turkish fighter jet was shot down by Syria over the Mediterranean in June 2012, after Syrian forces said it had entered the country's airspace.
Ours	A Turkish air force pilot has been killed after his jet crashed near the Syrian border , officials say.	

Table 2: Sample summaries from PEGASUS (Base) and EFACTSUM (Ours) on XSUM articles. The information from the article that contradicts the Base summaries is in **bold**. We can see that the outputs from our fine-tuned model not only generate faithful summaries but also capture the essential information from the article well.

Figure 2: Extractiveness scores for the various models on both the datasets. The x-axis represents the Extractiveness calculated using the *coverage* score defined by Grusky et al. (2018). Smaller the extractivness score, more the abstraction in the summaries.

Dataset	Model	MINT	μFactCC
CNN/DM	BART	57.94	42.14
	CLIFF	52.18	39.77
	EFACTSUM	**60.70**	**47.47**
XSUM	PEGASUS	25.21	44.12
	CLIFF	25.31	43.36
	EFACTSUM	**31.24**	**48.61**

Table 3: Abstractiveness scores as calculated by Dreyer et al. (2023) and abstractiveness-adjusted FactCC.

also presented in Appx. §B. We can observe that the extractiveness score for our model (EFACTSUM) is lesser than other models; it also achieves higher MINT scores (Tab. 3), which measures the abstractiveness of the summaries. Additionally, EFACTSUM shows higher scores for abstractiveness calibrated FactCC metric (μFactCC) for both datasets. This clarifies that the additional gains in factuality are not at a cost of absractiveness.

3.4 Ablation Study

In order to justify the modification made in the candidate ranking process of EFACTSUM, we compute baselines that highlight the importance of each individual component. We perform the following studies using PEGASUS fine-tuned on XSUM.

Candidate Selecting Process. As explained in §2 we restrict the number of candidates summaries in-order to maintain a class *balanced* set. We relax this constraint by simply scoring *all* the candidate summaries using FactCC. This is represented by EFACTSUM- w/o select. in Tab. 4. We can observe that this process leads to improved model factuality but still falls far short of the main approach by 4 points. Hence highlighting the advantage of focusing on generating quality training data.

Dual Scoring Technique. To understand the importance of using ROUGE to select the top candidates from both factual and non-factual sets, we ablate this step by selecting the top factual and non-factual summaries using FactCC itself. This is marked as EFACTSUM- w/o ROUGE in Tab. 4. Although the gains from this model on factuality are almost the same as EFACTSUM, it negatively affects the ROUGE score.

Model	R-L	FactCC
PEGASUS	39.26	24.33
EFACTSUM- w/o select.	38.32	26.38
EFACTSUM- w/o ROUGE	38.34	29.83
EFACTSUM	**39.45**	**30.48**

Table 4: Evaluation results for the various baseline models in §3.4. We can observe that both the components in the ranking strategy is required in order to obtain maximum benefits from the training process.

4 Related Work

Factual consistency in abstractive summarization has garnered much attention recently (Goyal and Durrett, 2020; Zhu et al., 2021). Existing works have explored improving factual consistency during fine-tuning, inference, and pre-training stages, respectively. For factual fine-tuning, works have applied contrastive learning (Cao and Wang, 2021; Nan et al., 2021b), reinforcement learning (Gunasekara et al., 2021) or knowledge integration (Zhu et al., 2021) to teach the model identify summaries of high factual consistency while Wan and Bansal (2022) modify the pretraining process to introduce factuality-awareness. Several works have also improved summary factuality through post-processing in inference, such as correcting errors and re-ranking by factual scores (Cao et al., 2020; Dong et al., 2020; Balachandran et al., 2022; Chen et al., 2021; Zhu et al., 2021). Our work differs from the aforementioned works as we improve both factuality and summary quality, unlike other methods, which often sacrifice one for the other.

5 Conclusion

We present EFACTSUM (**E**ffective **Fact**ual **Sum**marization), a candidate summary generation and ranking technique for contrastive summarization training, which helps make models more faithful without adversely affecting summary quality. Results show that this simple, yet effective method can achieve consistent gains on both factuality and similarity-based metrics without negatively affecting the degree of abstractiveness. We hope that our findings will encourage future research on factuality-consistent summarization to focus more on the tradeoffs between summary quality and factuality.

Acknowledgement

We appreciate the reviewers for their insightful comments and suggestions. We would also like to thank Raj Dabre and Sumanth Doddapaneni for their feedback on the initial versions of the work. Tanay Dixit was supported by the NSF REU Site Grant 2051101. Fei Wang was supported by the Annenberg Fellowship at USC. Muhao Chen was supported by the NSF Grant IIS 2105329, by Air Force Research Laboratory under agreement number FA8750-20-2-10002, by an Amazon Research Award and a Cisco Research Award. Computing of this work was partly supported by a subaward of NSF Cloudbank 1925001 through UCSD.

Limitations

While our approach helps train factuality-aware summarization models, it comes at an additional computation cost. It takes 3X time to train compared to the vanilla cross-entropy model. There is also an additional overhead computational cost in generating and scoring the candidate summaries for each article in the training dataset, but we believe that the gains justify the additional computation cost. Improving faithfulness in summarization models is a challenging task. Although we make improvements over prior work by achieving improved factuality metrics, like the compared prior works, our work has not focused on numerical consistency. This could be a meaningful research direction for follow-up work.

References

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

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.

Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, and Alexander Gray. 2022. X-FACTOR: A cross-metric evaluation of factual correctness in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7100–7110, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941.

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.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2023. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

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.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Ranit Aharonov, and Sachindra Joshi. 2021. Using question answering rewards to improve abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 518–526, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

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.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

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.

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.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021a. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021b. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894, Online. Association for Computational Linguistics.

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.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

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.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

A Additional Training Details

All experiments were carried out using 4, 24GB NVIDIA RTX A5000 GPUs. Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 kgCO_2eq/kWh. Total emissions are estimated to be 4.84 kgCO_2eq of which 0 percents were directly offset. Estimations were conducted using the MachineLearning Impact calculator presented in (Lacoste et al., 2019).

XSUM : For every news article in XSUM, we use diverse beam search (Vijayakumar et al., 2018) to generate 16 summaries using fine-tuned PEGASUS[3] and 16 summaries using CLIFF (*maskrel, syslowcon, swapent* and *regenrel*). We use the standard ROUGE-L[4] implementation and for FactCC, we use the `FactCC` checkpoint from the official implementation provided by the authors[5]. Articles for which we are unable to generate the required number of factual and non-factual summaries are discarded. In the end, our training dataset contains 145,040 data points. Choosing a bigger candidate size (>6) led to a decrease in the training dataset size as mentioned in §2.

Hyperparameters	Value
model	google/pegasus-xsum
no. of params	568M
max learning rate	1e-4
warmup steps	500
number of epochs	5
per device batch size	1
accumulation step	16
margin	0.001
max seq length	512
mle weight	1
ranking weight	10

Table 5: Hyperparameters for PEGASUS on XSUM.

CNN/DM For CNN/DM we follow the same process as described for XSUM, except here we use BART Large[6]. For CLIFF on CNN/DM we use *syslowcon_maskrel, syslowcon, syslowcon_swapent* and *syslowcon_regenrel* models. In the end our training dataset has 246,796 articles.

[3]google/pegasus-xsum

[4]https://github.com/summanlp/evaluation/tree/master/ROUGE-RELEASE-1.5.5

[5]https://github.com/salesforce/factCC

[6]facebook/bart-large-cnn

Training details For training we use the Adam optimizer with linear learning rate scheduling for the model training. Tab. 5 and Tab. 6 contain the best set of hyper-parameters for training PEGASUS on XSUM and BART on CNN/DM. These hyper-parameters were obtained after an extensive grid search. We perform validation after every 1600 steps and save the best model using the validation cross-entropy loss.

Hyperparameters	Value
model	facebook/bart-large-cnn
no. of params	400M
max learning rate	3e-5
warmup steps	500
number of epochs	5
per device batch size	1
accumulation step	16
margin	0.001
max seq length	1024
mle weight	0.1
ranking weight	10

Table 6: Hyperparameters for BART on CNN/DM.

Decoding parameters We follow Cao and Wang (2021) and use the beam search algorithm to decode summaries. For BART, we set the beam sizes as 4 on CNN/DM and a beam size of 8 is used for PEGASUS on XSUM. The additional decoding parameters are in Tab. 7.

Hyperparameters	Value
BART	
beam size	4
length penalty	2
max-length	140
min-length	55
PEGASUS	
beam size	8
length penalty	0.6
max-length	62
min-length	11

Table 7: Decoding parameters for BART and PEGASUS

B Extractiveness Results

The extractivenes scores as calculated using the *coverage* score defined by Grusky et al. (2018) are present in Tab. 9 and Tab. 8. Lower the score the higher the abstraction. We can observe that EFACTSUM achieves a lower abstraction level than CLIFF on both the datasets.

Model	Abstractiveness (↓)
Reference	**0.666**
Pegasus	0.735
CLIFF	0.759
EFACTSUM	0.720

Table 8: Extractivness analysis for XSUM

Model	Abstractiveness (↓)
Reference	**0.880**
BART	0.991
CLIFF	0.989
EFACTSUM	0.979

Table 9: Extractivness analysis for CNN/DM

C Generated outputs

More examples generated outputs by EFACTSUM on different backbones and raw documents are in Tabs. 10 and 11.

System	Summary	Article
Base.	The number of migrants and refugees arriving on the Greek island of Lesbos has halved in the past week.	Lesbos used to get more than 5,000 a day. On Monday there were just four. But with Europe's borders closed, more than 50,000 migrants remain in Greece waiting for a decision about their futures. . . . But here she is in Moria, once a transit camp for migrants, now **since the EU deal with Turkey**, a detention centre, run by central government. . . . It is another sign of how Greece was simply overwhelmed by the numbers who came, while itself in the middle of an economic crisis. Most of those who arrived before March 20, the start of the EU-Turkey agreement, are free to come and go, but cannot leave the island. Those who came after that date are locked in, waiting for a decision . . .
Ours	The number of migrants arriving on the Greek island of Lesbos has halved since the EU struck a deal with Turkey to stem the flow .	
Base	Hundreds of eggs from two rare bird species have been stolen.	The Mediterranean gull and black-headed gull eggs were illegally harvested from from islands in Poole Harbour, Dorset. . . . Natural England is urging any restaurants or pubs to ask to see a valid licence before buying eggs to prepare in meals. Birds of Poole Harbour had been surveying a group of islands in the harbour when the theft was discovered. Mediterranean gulls are classified as a Schedule One species, meaning anyone disturbing their nests must have a special licence. Paul Morton, who runs the charity, said Mediterranean gulls' eggs were not approved for human consumption, and could be a "health issue". "I'm distraught, really. To see the taking of hundreds and hundreds of eggs from an important colony is quite sickening," he said. Mr Moreton said there had been previous convictions for egg poaching in the last 10 or 15 years...
Ours	Hundreds of gull eggs have been stolen from a protected colony.	
Base	A volcano in western Indonesia has erupted for the second time in two years, killing at least 11 people, officials say.	The victims were farming in an area that was declared unsafe because of its close proximity to Mount Sinabung. The volcano was still spewing ash on Sunday, hampering rescue operations. More than a dozen people were killed when it erupted in 2014. It also erupted in 2010, after having been dormant for 400 years. Rescue teams are still scouring the area, looking for more victims who may have been killed or badly burned by the hot gas and ash clouds released in the eruption. Rescue teams were searching homes and farms in the village of Gamber, which was also evacuated in 2014. What causes volcanoes? The 2,460-metre (8,070 foot) tall volcano is among the country's most active. Indonesia, located on the Pacific Ring of Fire, has more than 120 active volcanoes.
Ours	At least 11 people have been killed after a volcano on the Indonesian island of Sumatra erupted , officials say .	
Base	The SNP and Labour have won seats on Edinburgh Council in two by-elections.	It was the first time the Single Transferable Vote (STV) system had been used to select two members in the same ward in a by-election. The SNP topped the vote in the Leith Walk by-election, while Scottish Labour won the second seat from the Greens. The by-election was called after Deidre Brock of the SNP and Maggie Chapman of the Scottish Greens stood down. . . . The turnout for the by-election was 25.1%. The SNP also held the Midlothian West seat on Midlothian Council with a swing of 6.3% from Labour. The party's Kelly Parry secured 1,540 votes, ahead of Labour's Ian Miller on 945 votes. The by-election was called after Owen Thompson was elected as SNP MP for the Midlothian constituency.
Ours	A by-election has been held in Edinburgh to fill two seats on the city council .	

Table 10: Sample summaries from PEGASUS (Base) and EFACTSUM (Ours) on XSUM articles.

System	Summary	Article
Base	Video shows the lions interacting with the visitors who stand inside a metal cage attached to a car. The video was captured by ekant veer, 35, an associate professor at the university of canterbury. A number of the lions are fed directly through the metal bars, while others receive meat dropped from the back of the cage.	visitors to a wildlife park in new zealand got to encounter a pride of lions up-close and personal. filmed at the orana wildlife park the countrys only open-range zoo the video shows the lions interacting with the visitors who stand inside a metal cage attached to a car. the video, which was captured by ekant veer, 35, an associate professor at the university of canterbury, also shows the lions scaling the cage and eating meat through its bars. . . . as the keeper speaks, the lion licks at a piece of meat that is held up against the bars as another lion walks across the roof of the cage. looking down at the people below, the lion wanders around as if deciding who it would like to make its prey before staring down the lens of the camera. set tongues wagging! one of the lions notices meat and begins sticking out its tongue in the hope of being fed. a lion stands next to one of the keepers and its large paw is the same size as the lady's head. the people inside can be seen recording the many lions from their phones, while another with paws the same size as the keepers head holds itself up against the cage and chews on some meat. later in the video people can be seen pointing out the various felines as a keeper moves her hand along the cage, instigating the lion to follow. still frames capture a lion standing up against the side of the cage alongside the keeper its power and size is plain to see. . . . orana wildlife trust. located on the outskirts of christchurch, the wildlife park is unique in that the people are caged in order to view the animals, not the other way around.
Ours	the video was filmed at the orana wildlife park in new zealand , the country 's only open-range zoo . the video shows the lions interacting with the visitors who stand inside a metal cage attached to a car . a number of the lions are fed directly through the metal bars , while others receive meat dropped from the back of the cage .	
Base	Taxpayers are having to find 11billion a year to top up the wages of millions of people working in supermarkets and other low paid jobs. Money is paid to some 5.2million workers in the form of tax credits and other benefits. Total amount of benefits paid to staff at some companies exceeds what the firms pay in corporation tax.	taxpayers are having to find 11billion a year to top up the wages of millions of people working in supermarkets and other low paid jobs. the money, which amounts to a massive public subsidy for the companies involved, is paid to some 5.2million workers in the form of tax credits and other benefits. . . . the charity is campaigning for the adoption of the living wage - 9.15 an hour in london and 7.85 for the rest of the uk - across both the public and private sector. it estimates this would reduce the need for in-work benefits by 6.7bn a year, which would make a massive dent in the 12billion reduction in welfare spending which the conservatives say is necessary. the current minimum wage for those over 21 is 6.50 an hour and will rise to 6.70 in october, da and sainsburys posted combined profits of 3.9bn last year, but between them cost the taxpayer more than 750m in benefits paid to their staff. tesco paid 519m in tax but received 364m in public subsidy for its 209,000 low-paid workers. asda spent 150m in tax but its 120,000 low-paid workers received 221m in benefits. . . . thesupermarkets said they paid above the minimum wage of 6.50 an hour for those aged over 21, regularly reviewed pay and gave employees benefits such as staff discounts. asda, which is part of the us retail goliath walmart, said pay and benefits should be considered in the round. in the usa, it is estimated that walmarts low-wage workers cost u.s. taxpayers an estimated $6.2 billion (4.2bn) in public assistance including food stamps, medicaid and subsidised housing. . . .
Ours	Taxpayers are having to find 11billion a year to top up the wages of millions of people working in supermarkets and other low paid jobs. Money is paid to some 5.2million workers in the form of tax credits and other benefits. Total amount of benefits paid to staff at some companies exceeds what the firms pay in corporation tax.	

Table 11: Sample summaries from BART Large (Base) and EFACTSUM (Ours) on CNN/DM articles.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3, Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.