X-FACTOR: A Cross-metric Evaluation of Factual Correctness in Abstractive Summarization

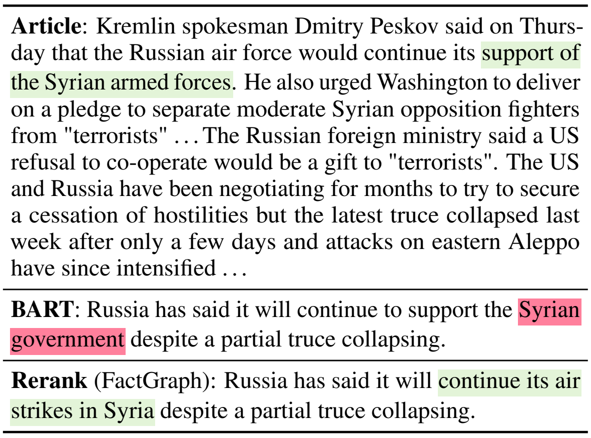
Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara,
Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan,
Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, Alexander Gray
{subhajit, sarath.swaminathan, chulaka.gunasekara, maxwell.crouse}@ibm.com
{srini, keerthiram.murugesan, ramon.astudillo, alexander.gray}@ibm.com
daiki@jp.ibm.com, tnaseem@us.ibm.com, kapanipa@us.ibm.com
IBM Research

Abstract

Abstractive summarization models often produce factually inconsistent summaries that are not supported by the original article. Recently, a number of fact-consistent evaluation techniques have been proposed to address this issue; however, a detailed analysis of how these metrics agree with one another has yet to be conducted. In this paper, we present X-FACTOR, a cross-evaluation of three high-performing fact-aware abstractive summarization methods. First, we show that summarization models are often fine-tuned on datasets that contain factually inconsistent summaries and propose a fact-aware filtering mechanism that improves the quality of training data and, consequently, the factuality of these models. Second, we propose a corrector module that can be used to improve the factual consistency of generated summaries. Third, we present a re-ranking technique that samples summary instances from the output distribution of a summarization model and reranks the sampled instances based on their factuality. Finally, we provide a detailed cross-metric agreement analysis that shows how tuning a model to output summaries based on a particular factuality metric influences factuality as determined by the other metrics. Our goal in this work is to facilitate research that improves the factuality and faithfulness of abstractive summarization models.

1 Introduction

In this work, we consider the task of automatic text summarization, i.e., the task of generating a concise summary of a given document that preserves its most salient information (Maybury, 1999). This is a prominent task that has been applied across a variety of contexts and domains, such as the summarization of legal documents (Kanapala et al., 2019), automatic news summarization (Fabbri et al., 2019), and summarization of dialog between agents and clients (Feigenblat et al., 2021).



Article: Kremlin spokesman Dmitry Peskov said on Thursday that the Russian air force would continue its support of the Syrian armed forces. He also urged Washington to deliver on a pledge to separate moderate Syrian opposition fighters from "terrorists" . . . The Russian foreign ministry said a US refusal to co-operate would be a gift to "terrorists". The US and Russia have been negotiating for months to try to secure a cessation of hostilities but the latest truce collapsed last week after only a few days and attacks on eastern Aleppo have since intensified . . .

BART: Russia has said it will continue to support the Syrian government despite a partial truce collapsing.

Rerank (FactGraph): Russia has said it will continue its air strikes in Syria despite a partial truce collapsing.

Figure 1: Factual hallucinations on an XSum article by BART (Lewis et al., 2019) summarization model which are not supported by the article. Here, we show the result of reranking, one of our factuality-aware methods, on this example that removes the hallucinations making the summary factually consistent with the article.

Most modern approaches to automatic text summarization can be characterized as either (a) *extractive* in that they select a subset of sentences from the longer source article to serve as the summary, or (b) *abstractive* in that they generate a summary that includes sentences and phrases not occurring in the source article. Abstractive summarization is generally considered more challenging, as the aspect of summaries containing information outside their source articles adds complications to both training and automatic evaluation. One of the most notable challenges faced when designing abstractive summarization models is their proclivity for generating factual errors (Cao et al., 2018; Maynez et al., 2020), i.e., for generating summaries that do not agree with the original document's text.

The fact that these models struggle with factuality is somewhat unsurprising, as the lexical overlap metrics (e.g., ROUGE (Lin and Hovy, 2003), BLEU (Papineni et al., 2002)) frequently used to evaluate and tune these models only compute the co-occurrence of n-grams between the reference

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7100–7110
December 7-11, 2022 ©2022 Association for Computational Linguistics

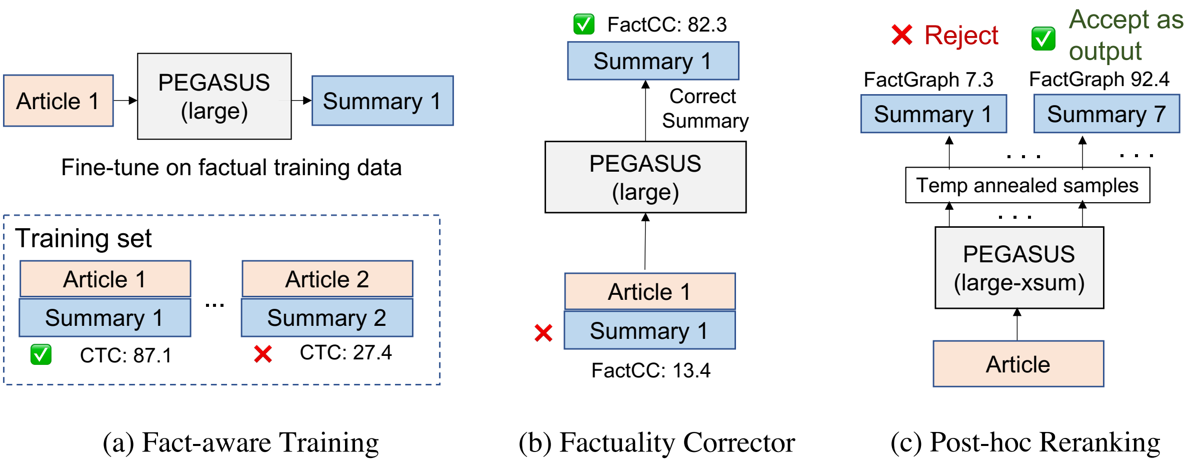


Figure 2: Three factuality-aware summarization methods that we use in this work for improving factual consistency.

and generated summary texts, and thus do not explicitly measure for factual correctness (Falke et al., 2019; Kryściński et al., 2019). In service of this issue, a large number of recent works have proposed metrics that are specifically targeted towards evaluating and improving the factuality of abstractive summarization (Zhu et al., 2021; Goyal and Durrett, 2021; Kryściński et al., 2020; Yuan et al., 2021; Ribeiro et al., 2022; Gunasekara et al., 2021). However, it is unclear how the factuality metrics introduced by these works agree with each other and how tuning summary generation with a particular metric influences how the summary is scored by other metrics. Therefore, in this work, we perform a more comprehensive and unified analysis of how these metrics can be used to improve factual consistency and how they interact with one another.

Our analysis is performed on seven of the recent factuality metrics. We leverage the following three different strategies for fact-aware summarization using these metrics: (a) Fact-aware training: a methodology to select factually consistent training data to fine-tune a summarization model; (b) Post-hoc reranking: Reranks the output summaries with respect to the factuality metrics; and (c) Fact-aware corrector: Leverages the scores from the metrics to train a seq2seq model that modifies a hallucinated summary to a factual text generation output. While previous works have performed some of the above fact-aware methods usually on a single metric, this work provides an extensive and unified analysis of these methods optimized for various factuality metrics. In this work, we are primarily interested to understand how optimizing fact-aware summarization methods for each of these metrics affects the performance of other lexical and factual evaluation metrics and how that can be used to design an abstractive summarization pipeline that outperforms recent fact-aware summarization baselines. Figure 1 gives an illustrative example.

Our contributions in this paper are as follows: (i) We present a comprehensive analysis of fact-aware summarization methods optimized for various factuality metrics and their impact on other metrics which has not been studied by previous works; (ii) From our analysis, we find that optimizing the method for lexical overlap metrics does not correlate well with factuality scores, whereas, optimizing for one of the factuality metrics can show gains for other factuality based metrics.

2 Fact-Aware Summarization

In this section, we detail the three methods we use to optimize for each of the factuality metrics and in turn for analyzing the cross-metric agreement. Figure 2 illustrates the three methods of Fact-aware Training (FactTr), Post-hoc reranking (rerank), and Hallucination corrector (corrector).

2.1 Fact-aware Training (FactTrain)

Abstractive summarization datasets such as XSum are known to contain hallucinations in the gold summaries (Maynez et al., 2020) that cause models fine-tuned on these datasets to produce factually inconsistent outputs. To combat this issue, various methods (Goyal and Durrett, 2021; Zhao et al., 2020; Cao et al., 2018) have been proposed, most often using post-hoc corrections to improve the quality of summarization. Here we leverage factuality metrics to improve the quality of the training data, rather than to correct the outputs of an abstractive summarization model.

Specifically, consider that we have a original training set of $\mathcal{D} = \{(x_i, y_i)\}_i$ and we compute a factuality score $s_i = g_s(x_i, y_i)$, where g_s denotes the scoring method. Given a threshold t, we produce a factually-consistent training subset by retaining the article-summary pairs that score above the threshold, $\mathcal{D}^s_{cl} = \{(x_i, y_i) \mid s_i > t\}_i$. While previous methods only used a single metric for this downstream task, we use multiple recent metrics for our factual training subset extraction. We computed the factuality scores using $g_s = \{\text{ANLI}, \text{SummaC}, \text{FactCC}\}$ methods on the XSum training and validation set and fine-tuned a pre-trained Pegasus-large (Zhang et al., 2020) model on each of those factually-consistent subsets $\mathcal{D}^s_{cl}$. The best model on the validation set was chosen as the final model for testing. We experimented on several threshold values and the best threshold was picked according to the performance on the validation set.

2.2 Post-hoc Reranking (Rerank)

Most often, hallucinations occur due to incorrect entities (or quantities) in the summary compared to the original articles. Previous works (Falke et al., 2019; Chen et al., 2021) have tried to correct these by ranking generated summaries according to classification model probabilities. We take a similar approach but on a much larger set of metrics for obtaining the reranking scores.

Since the factuality metrics are not reference-based (do not require reference summaries) we can compute the factuality score on each example of the test set. Given an article x, our re-ranking method first generates N sample summaries from the Pegasus model $f(.)$ at various temperatures T_k such that $\hat{y}_k = f(x; T_k)$. Next, we obtain the factuality scores and then choose the best summary $\hat{y}_{k^*_s}$, where the index of best summary is $k^*_s = \arg\max_k g_s(x, \hat{y}_k)$ for the factuality metric s. We obtain the best summary for each example and each metric (listed in Section 3.3) and also compute the cross-metric scores for each such example on the FactCollect test set.

2.3 Hallucination Corrector (Corrector)

Learning a model to improve factuality is another approach for factually-aware summarization generation. Wan and Bansal (2022) recently proposed a corrector module with masked fine-tuning to correct hallucinations in the summary output. Following the success of model-based factuality evaluation methods (Ribeiro et al., 2022; Kryściński et al., 2020), we use a model-based factuality improvement approach in this work. We fine-tune a corrector model to predict a factually consistent summary from the hallucinated summary and corresponding article together as input.

To generate hallucinated and factual summaries from the training set, we use the temperature annealed sampling similarly to the `rerank` method to generate a spectrum of factuality scores on the generated summaries. Given a article x from the training set, we get $N = 10$ summary sample using Pegasus model fine-tuned on Xsum $f(.)$ at various temperatures and compute the scores as $\mathcal{M}_s = \{g_s(x, f(x; T_k))\}_k$. We then pick the (worst summary | article) and best summary as the input-output pair and add that to the training set given the best and worst scores differ by some percentage threshold. Following this, we fine-tune a pre-trained PEGASUS-large model on this data to predict the factual summary as output. After training, we use this to improve the factual consistency of the outputs PEGASUS fine-tuned model trained on the full Xsum dataset.

3 Experimental results

In this section, we outline the dataset we used, our evaluation metrics, and experimental results.

3.1 Dataset

We conducted our experiments on the XSum dataset (Narayan et al., 2018) that is well suited for abstractive summarization settings. This dataset consists of articles from the British Broadcasting Communication (BBC) and a one-sentence summary of the article. Maynez et al. (2020) report that more than 70% of the gold summaries in this dataset have "hallucinations" and hence this dataset is a good candidate for studying factuality. We also report results on the CNN/DailyMail dataset (Nallapati et al., 2016) to show that our analysis generalizes across domains. For testing, we use a subset of this dataset contained in the FactCollect dataset (Ribeiro et al., 2022) that is comprised of four human annotated factuality datasets (Kryściński et al., 2020; Wang et al., 2020; Maynez et al., 2020; Pagnoni et al., 2021) consisting of samples from Xsum and CNN/DailyMail datasets that are used to benchmark factuality metrics.

Method	Lexical Overlap			Factuality						
	R1	R2	RL	ANLI	SummaC	Q2	BartScore	CTC	FactCC	FactGraph
Pegasus	44.1	22.2	38.0	75.9	22.9	7.06	6.3	78.6	32.62	19.3
UniLM (FC)	41.9	19.9	34.2	54.7	21.6	4.9	4.1	77.5	27.6	12.1
FASUM	30.8	11.0	24.2	20.8	21.0	2.4	3.4	73.4	35.1	3.9
TConv2S (FC)	31.7	12.0	26.6	33.2	21.2	2.4	3.5	-	31.9	5.4
CLIFF	**44.5**	**22.7**	37.1	65.9	21.5	5.1	6.2	79.6	31.0	16.6
FactTrain (ANLI)	43.6	21.6	36.9	69.3	23.1	4.8	6.5	79.4	35.4	18.6
FactTrain (FactCC)	42.0	19.6	35.1	66.7	23.2	5.5	7.5	80.7	44.5	24.0
FactTrain (SummaC)	38.1	17.1	31.2	69.7	27.3	-	**9.0**	**81.6**	**38.8**	**30.6**
Corrector (ANLI)	44.3	22.3	38.2	67.3	23.4	5.3	6.0	79.0	34.7	22.1
Corrector (FactCC)	44.4	22.3	38.3	66.5	23.4	5.3	5.9	78.8	35.5	22.6
Corrector (SummaC)	43.8	21.9	37.6	71.3	**23.8**	5.5	6.3	79.3	34.5	25.1
Rerank (ANLI)	44.0	21.9	38.2	86.8	23.4	6.6	6.3	79.3	32.0	20.1
Rerank (SummaC)	44.0	22.3	38.4	77.6	24.1	6.7	6.3	78.4	33.4	21.1
Rerank (Q2)	43.9	22.4	37.9	77.0	23.2	10.0	6.4	79.1	32.0	18.3
Rerank (BartScore)	**44.5**	**22.7**	**38.4**	77.5	23.2	7.4	7.6	79.8	32.2	21.2
Rerank (CTC)	44.3	22.5	38.3	**78.5**	23.3	**7.5**	6.8	81.9	29.5	19.6
Rerank (FactCC)	43.6	21.6	37.1	76.0	23.0	7.00	6.3	78.5	45.1	20.0
Rerank (FactGraph)	43.7	21.9	37.7	75.5	23.4	7.1	6.4	79.2	36.9	25.2

Table 1: Comprehensive analysis of the performance of fact-aware summarization methods on the Xsum samples from the FactCollect test set optimized for various factuality metrics, as measured by other factuality scores. Our methods can outperform other recent fact-aware methods like CLIFF on a number of factuality metrics.

	Lexical overlap			Factuality-based metrics			
Method	R1	R2	RL	SC	CTC	FCC	FG
Pegasus	**38.8**	**19.1**	**29.1**	38.2	87.0	42.5	78.0
SC	35.9	17.8	28.0	52.5	**89.1**	**50.9**	79.5
CTC	36.0	17.8	28.2	**44.7**	91.5	50.6	**82.3**
FCC	36.6	18.4	28.3	42.4	88.1	65.1	80.5
FG	35.7	17.8	27.9	44.0	88.5	49.1	88.6

Table 2: Analysis of the performance of the re-ranking method on the CNN DailyMail samples from FactCollect test set optimized for various factuality metrics. The results show that factuality metrics improve when the reranking model optimizes for factuality metrics. SC, FCC, and FG are SummaC, FactCC, and FactGraph respectively.

3.2 Baselines

The following baselines were used to compare the results obtained by our models.[1]

UniLM: UniLM (Dong et al., 2019) is a pre-trained language model, which can be fine-tuned to many natural language generation tasks. The model is pre-trained using unidirectional, bidirectional, and sequence-to-sequence language modeling tasks. We fine-tune the UniLM model on the Summarization task.

TConv2S: In TConv2S (Narayan et al., 2018), the authors propose an abstractive summarization model which is conditioned on the article's topics and implemented using convolutional neural networks (CNN). The model uses the topic distributions obtained from Latent Dirichlet Allocation (Blei et al., 2003) in the summarization network as additional input.

FASum: FASum (Zhu et al., 2021) proposes to extract and integrate factual relations represented in a graph into the summary generation process. The model uses a graph attention network (Veličković et al., 2018) to obtain the representation of each node, and fuse that into a transformer-based encoder-decoder architecture via attention.

CLIFF: CLIFF (Cao and Wang, 2021) proposes a contrastive learning (CL) based training objective that drives summarization models to expand the margin between factually consistent summaries and their incorrect peers. Once the representations are learned including the CL-based objective, the transformer-based seq2seq models are fine-tuned to

[1] We also considered comparing with FactPegasus (Wan and Bansal, 2022). However, the pre-trained models for FactPegasus were not released during this work.

generate the summaries. We use the SYSLOWCON model from Cao and Wang (2021) as the baseline.

3.3 Evaluation metrics for Factuality

In this section, we summarize the evaluation metrics that were selected from the literature to evaluate the factuality of the models.

ROUGE: ROUGE (Lin, 2004) (Recall-Oriented Understudy for Gisting Evaluation) is a common metric for evaluating text generation tasks including abstractive summarization. ROUGE measures the overlapping text in terms of n-grams and word sequences between the gold summary and the model outputs.

ANLI: ANLI (Nie et al., 2019) (Adversarial Natural Language Inference) is a prominent textual entailment dataset. Re-ranking summaries based on probabilities from models trained on ANLI have shown to be effective in improving the factual correctness of the summaries (Barrantes et al., 2020). Honovich et al. (2022) has shown that models trained on ANLI datasets show good classification performance for factuality detection. Therefore, we use ANLI as one of our factuality metrics.

SummaC: SummaC (Summary Consistency) (Laban et al., 2021) is related to the previous metric (ANLI). SummaC addresses the issue with granularity in NLI models (sentence vs document) used for inconsistency detection of summarization. The scores from SummaC reflect inconsistencies of summary sentences occurring in any place in the source article.

Q2: Q^2 (Honovich et al., 2021) follows the Question Answering and NLI-based metric, however, with an inclusion of a knowledge source against which the factual consistency is measured. Honovich et al. (2022) showed that this metric can be reliably used for factuality detection in abstractive summarization tasks and hence, we use it as one of our metrics.

BARTScore: BartScore (Yuan et al., 2021) is built upon BART (Lewis et al., 2019), a pretrained encoder-decoder architecture to evaluate text generation models. The scoring mechanism can be used in an unsupervised fashion where the most used metric is the weighted log probability of gold truth given the generated text.

CTC: Compression Transduction Creation (CTC) (Deng et al., 2021) proposes a unifying perspective based on the nature of NLG tasks, including compression (e.g., summarization), transduction (e.g., text rewriting), and creation (e.g., dialog). Information alignment, which is defined as the overlap between the input, context, and output, is critical for all three of the above categories. CTC metric adopts contextualized language models to measure the information alignment, specifically for consistency and relevance which are necessary components for compression (summarization).

Model	RL	CTC	FactCC	FactGraph
Pegasus	38.0	78.6	32.62	19.3
Rerank (FactGraph)	37.7	79.2	36.9	25.2
Rerank + Corrector (ANLI)	38.3	79.1	37.7	25.7
Rerank + Corrector (FactCC)	**38.4**	79.2	**37.9**	27.1
Rerank + Corrector (SummaC)	38.0	**79.7**	37.7	**28.9**
FactTr (FactCC)	35.1	80.7	**44.5**	24.0
FactTr + Corrector (ANLI)	35.3	80.8	43.9	27.4
FactTr + Corrector (FactCC)	35.5	80.6	**44.5**	26.8
FactTr + Corrector (SummaC)	**35.6**	**80.9**	40.6	**28.8**

Table 3: Ablation results for our fact-aware summarization methods. The combination of corrector with Rerank+Corrector and FactTrain+Corrector methods improves both n-gram based and factuality metrics.

FactCC: FactCC (Kryściński et al., 2020) is a BERT-based classification model to determine where a given text/summary is CONSISTENT/INCONSISTENT with the source article. The model is trained on synthetic data generated by transforming ground truth with paraphrasing, swapping entities, numbers, pronouns, etc.

FactGraph: FactGraph (Ribeiro et al., 2022) uses both text and its semantic graph representations (AMR) to enhance the factuality of the summaries with respect to the source article. Similar to FASum (Zhu et al., 2021), FactGraph jointly uses text encoder and graph convolutions as graph encoder, and the model outputs a factuality score.

3.4 Fact-aware summarization performance

Our main results are shown in Table 1 which measure the performance of various methods optimized for seven factuality metrics and also measured across those metrics. For the baseline methods, we use the summarization outputs provided from FASUM (Zhu et al., 2021) and (Cao and Wang, 2021). For the corrector results, we picked the output of fine-tuned Pegasus model and used that as the input for the corrector module to generate corrected summaries. For this comparison, we ignored the diagonal entries in the rerank section since those will be biased because the same metric was used for reranking in these cases. Similarly matching optimizing and measuring metrics for FactTrain is also not considered the best results.

In the baseline method, CLIFF (Cao and Wang, 2021) is one of the recently proposed fact-aware methods that optimize both ROUGE and factuality scores. As a result, CLIFF obtains a high score for the n-gram metrics and improves on some factual metrics. In comparison, our methods show better performance on the factuality metrics like BartScore, CTC, FactGraph, and FactCC, with a small sacrifice on the ROUGE metrics scores. This might be due to the fact that the Xsum dataset has almost 70% hallucinations (Maynez et al., 2020) in the gold summaries. Therefore optimizing factuality might generate summaries that do not match the original hallucinated gold summary, resulting in a low ROUGE score. Additionally, FactTrain might reduce the number of available training samples. Therefore the model is trained on such limited data which might lead to low ROUGE score performance.

Specifically, FactTrain trained with SummaC metric shows the best performance on multiple factuality metrics. On average, our FactTrain method got lower ROUGE scores while getting a better score for factual consistency. The Rerank method gets better ROUGE scores, while also improving the factual consistency metrics by a small amount. This is because the rerank method uses Pegasus trained on the original Xsum dataset for generating the samples which inherently have hallucinations but have good lexical overlap with the reference summary. We observe reranking can only remove certain entity-level hallucinations and hence is limited in its capacity to improve summarization. The corrector module shows better scores compared to the baseline with no significant drop in lexical overlap performance.

Table 2 shows the various factuality scores by various metrics (columns) using the reranking method optimized for various factuality metrics (rows) for CNN/DailyMail samples. Similar to the Xsum samples from the FactCollect dataset, CNN/DailyMail samples also shows that optimizing for lexical overlap metrics does not improve the factuality scores and vice-versa. This result shows that our analysis generalizes across multiple domains.

3.5 Ablation study

Table 3 shows the results of our ablation study with the combination of various fact-aware summarization methods. We show the results with Rerank+Corrector and FactTrain+Corrector setting for three metrics - ANLI, FactCC, and SummaC. The corrector module improves the factuality in both cases. SummaC based corrector module gives the best overall improvements with FactGraph scores of 28.9 and 28.8. These results show evidence that the combination of our fact-aware summarization methods shows improvements over the individual modules.

3.6 Effect of thresholds for FactTrain

Table 4 shows the result on how changing the threshold for filtering the dataset affects the ROUGE-L score and the factuality scores. The factuality scores improve with increasing the threshold for both cases. As the threshold increases, the model is trained with factually consistent article, summary pairs hence improving the factuality scores. However, the number of training points decreases with an increasing threshold. For example, the number of training points for ANLI are 6925, 6476, and 5906 for 0.9, 0.95, and 0.98 respectively. Due to this phenomenon, the model is fine-tuned on a lesser number of points which might explain the decreasing ROUGE-L scores in this case.

3.7 Correlation Between the Metrics

Table 5 shows the Pearson correlation between the metrics as measured on the FactCollect test set (Xsum samples). We use the reranking method's dataset to compute these correlations. For each metric (on each row), we obtain the best scores from 10 samples of all the data points and compute the correlation with the other metrics by collecting those other scores for the best summary indices. The diagonal entries where the reranking and cross-analysis metrics are the same should be ignored. Our analysis shows that the factuality metrics show a good correlation with each other whereas shows a poor correlation with lexical overlap metrics. For example, the correlation between FactGraph and R1 is -0.155, showing a negative correlation due to the factual inconsistencies in the Xsum gold summaries. This phenomenon of inverse correlation between FactGraph and ROUGE scores has also been seen in previous experiments in Table 1. On the other hand, the metrics SummaC and FactGraph show a higher correlation of 0.389 possibly depicting strong inter-metric agreement.

Table 6 shows the Pearson correlation between metrics from the CNN/DailyMail samples from the

FactCollect test set. Similar to the above results, in this case factuality metrics exhibit negative/low correlation with lexical overlap metrics but considerably higher correlation with other factuality metrics. This study gives us an insight into how the different factuality metrics agree with itself which can be useful for downstream tasks of factuality improvements and this analysis holds across multiple datasets.

3.8 Correlation with human judgements

We present results on how each of the metrics compares with human-annotated factuality labels. We use AUROC score to measure the agreement of the metrics with human labels inspired from Honovich et al. (2022). We report the scores on the full FactCollect test set for the following metrics, ANLI: 0.81, R1: 0.38, R2: 0.42, RL: 0.37, SummaC: 0.92, CTC: 0.92, Q2: 0.83, FactCC: 0.84, FactGraph: 0.95. The results show that factuality metrics specially FactGraph show better agreement with human-annotated factuality labels. It is not surprising that FactGraph shows the best AUROC because it was trained on the FactCollect train set. Lexical overlap metrics show poor agreement with human-annotated factuality labels.

3.9 Qualitative analysis

Table 7 shows an example of hallucinations in the Xsum dataset. The article is about a potential crisis in South Sudan. The output from Pegasus hallucinates that the war is described as “world’s worse since World War Two”, however, there is no such evidence in the original article. Even the gold summary from the dataset has a hallucination of “three years” that is not found in the article. We show the results of our methods’ outputs for FactTrain with ANLI and SummaC, which produces summaries that are very different from the gold article but is factually consistent with the article. The SummaC output mentions “UNHCR” and “genocide and ethnic cleansing” which is supported by the article as highlighted with green color. For ANLI, the summary produced is factual but does give much detail about the article’s content which can occur due to primarily optimizing for a factuality-based metric. A potential solution to this can be using a weighted combination of lexical overlap and factuality metrics. We also show the output from FactCC-based reranking output that produces a similar summary as FactTrain (SummaC) but removes the information about the subject of the sentence.

	Threshold	RL	CTC	FactCC	FactGraph
ANLI	0.9	**37.9**	79.7	34.5	16.2
	0.95	36.9	79.4	**35.4**	**18.6**
	0.98	37.1	**79.8**	34.9	17.9
FactCC	0.9	34.4	80.0	45.1	17.6
	0.95	**35.5**	80.2	**47.9**	23.5
	0.98	35.1	**80.7**	44.5	**24.0**

Table 4: Effect of threshold on creating the fact-aware training subset for ANLI and FactCC. Increasing the threshold yields better factuality scores with some drop in performance for RougeL (RL) score.

4 Related Works

Abstractive summarization has been an important task in evaluating models’ abilities to understand language (Lewis et al., 2019; Zhang et al., 2020). Recently, multiple studies have shown that the quality of the summaries being generated has issues with factuality and being faithful to the source article. In other words, the recent summarization models hallucinate content that is mostly erroneous (almost 70%) in the context of the source article (Cao et al., 2018; Falke et al., 2019; Maynez et al., 2020). Such insights into the factual content in the summaries of models have opened up research efforts both in terms of developing models (Wan and Bansal, 2022; Kang and Hashimoto, 2020) and improving existing evaluation metrics to capture factuality and faithfulness as a measure to improve the models (Ribeiro et al., 2022; Kryściński et al., 2019; Xie et al., 2021). SummEval (Fabbri et al., 2021) also study re-evaluation of factuality scores for abstractive summarization. Models and approaches in improving factuality and faithfulness in summarization can be categorized into (a) Post-hoc reranking-based approaches: top-k summaries (modified) from the existing models are re-ranked based on models that ensure factuality and faithfulness (Chen et al., 2021; Falke et al., 2019; Dong et al., 2020; Zhao et al., 2020), (b) Graph-based methods: these set of methods not only leverage text of the article and summary, but also their graph representations from dependency and semantic parses such as OpenIE and AMR, and (c) Loss-based methods. Below, we go into detail about each of these works in comparison to ours.

Post-hoc re-ranking or correction: Post-hoc re-ranking based approaches mostly have two stages. The first is a candidates selection phase, where either top-k summaries from the summarization models are used or only the top summary

	Lexical overlap			Factuality-based metrics						
Method	R1	R2	RL	ANLI	SummaC	Q2	BartScore	CTC	FactCC	FactGraph
ANLI	0.155*	0.136*	0.118	1.0*	0.105	0.096	0.133*	0.173*	0.038	0.138*
SummaC	-0.01	0.008	0.0	0.144*	1.0*	0.287*	0.346*	0.291*	0.258*	0.389*
Q2	0.017	0.041	0.028	0.121	0.245*	1.0*	0.262*	0.165*	0.064	0.238*
BartScore	0.211*	0.163*	0.179*	0.181*	0.321*	0.209*	1.0*	0.235*	0.086	0.396*
CTC	0.094	0.14*	0.118	0.186*	0.286*	0.186*	0.224*	1.0*	0.025	0.236*
FactCC	-0.175*	-0.112	-0.14*	0.023	0.173*	0.057	0.036	0.015	1.0*	0.287*
FactGraph	-0.155*	-0.13*	-0.156*	0.201*	0.337*	0.151	0.418	0.22*	0.381*	1.0*

Table 5: Pearson correlation between the various metrics on the FactCollect dataset. Values marked with * are statistically significant with $p < 0.05$. The factuality metrics show a considerable positive correlation amongst each other while showing a negative/low correlation with the ROUGE score.

	Lexical overlap			Factuality-based metrics			
Method	R1	R2	RL	SC	CTC	FCC	FG
CTC	-0.04	0.08	0.04	0.48*	1.0*	0.45	0.59*
FCC	0.07	0.08	0.10*	0.34*	0.40*	1.0*	0.40
FG	0.02	0.10	0.08	0.24*	0.35*	0.37*	1.0*

Table 6: Pearson correlation between metrics on CNN Dailymail samples from the FactCollect dataset. SC is SummaC, FCC is FactCC and FG is FactGraph. Values marked with * are statistically significant with $p < 0.05$.

from the summarization model is used to generate candidates by replacing entities (or quantities) of the same type mentioned in the original article. These candidate summaries are them re-ranked using multiple different models such as (a) Classification models such as NLI model (Falke et al., 2019) and BART + Linear layers (Chen et al., 2021); (b) Span prediction (Dong et al., 2020) or sequence labeling models (Zhao et al., 2020). In this category of approaches, Falke et al. (2019) and Zhao et al. (2020) do not show any significant improvements in correcting factual errors, whereas the ones that generate candidates based on entities mentioned in the original article do have improvement with different factuality measures, with negative or no impact on the ROUGE scores.

Graph-based methods: This category of approaches includes not only the text of the article and the summary but a graph representation of the text using dependency parses and semantic parses such as OpenIE (Cao et al., 2018; Song et al., 2020) and Abstract Meaning Representation (AMR) (Liu et al., 2015). Specifically Liu et al. (2015) transforms the text summarization to graph summarization where the models are trained using AMRs (the graph structure) of the source article and the summaries. FTSum (Cao et al., 2018) uses triples extracted from text using OpenIE and dependency parser as additional information for summarization. The model includes a dual encoder that encodes text and triples with the decoder using an attention mechanism between them to output the summary. OpenIE is also used by FASum (Song et al., 2020), where the extracted triples are used to form a graph structure that is encoded using Graph Attention Networks. The model is built using a transformer where the decoder computes cross-attention over the nodes' embeddings from the knowledge graph in conjunction with the cross-attention over the encoder's embeddings. This methodology also includes a post-hoc fact corrected trained on synthetically generated data with entity replacement similar to post-hoc approaches (Dong et al., 2020).

Loss-based methods: The final category of methods focuses on addressing the problem of factuality by modifying the loss functions for training the model. Loss Truncation (Kang and Hashimoto, 2020) introduces distinguishability by introducing a surrogate loss based on negative entropy where the samples with the highest log loss will be removed. QUALS (Nan et al., 2021) uses a combination of summarization and question-answering methods for producing a factually consistent summary. The model back propagates contrastive loss through both the summaries and qa pairs. Constrained Abstractive Summarization (CAS) (Mao et al., 2020) introduces constrained decoding where the constraints are either manually or automatically added (key-phrases vocab from the source article). Our Fact-aware training which preprocesses and filters the training data is inspired by the Loss Truncation (Kang and Hashimoto, 2020) approach where the loss truncation learns to ignore training data with less distinguishability.

5 Conclusion

In this work, we provided a comprehensive analysis of the recent factuality metrics for abstractive sum-

Article: United Nations officials rarely use the words "genocide" and "ethnic cleansing" but they now say potentially both could envelop the world's youngest country. Since violence flared in Juba in July and spread to the previously peaceful southern Equatoria states of South Sudan, . . . "This has been unrelenting since July," said Nasir Abel Fernandes, the UNHCR's senior emergency coordinator in northern Uganda. "The international community has to pay attention, and pressure the South Sudanese leaders to stop this, as it's a massacre of civilians from both sides . . . Sudan is doing to its own people . . .
Pegasus : South Sudan's civil war could be the world's worst since World War Two, the UN has warned
Gold summary: For three years South Sudan has tumbled deeper into self-inflicted chaos, and it now finds itself on the brink of something even more terrifying.
FactTrain (ANLI): South Sudan's civil war is spiralling out of control.
FactTrain (SC): The UN refugee agency (UNHCR) has warned that South Sudan's civil war could lead to "genocide" and "ethnic cleansing"
Rerank (FactCC): South Sudan's civil war has been described as "genocide" and "ethnic cleansing"

Table 7: Factually-aware summarization using various methods studied in this work on an article from XSum dataset. The output from Pegasus shows factual hallucinations in this example and the gold summary also has hallucinated content. The summaries produced by our methods in this work produce summaries that are supported by the original article.

marization, particularly, to understand their impact on each other. The analysis has been driven by different strategies for building factually-aware summarization models. We present three fact-aware summarization methods in this work and showed that a simple methodology to optimize for factuality metrics can outperform existing strong baselines for a fact-aware summary generation. Furthermore, we have seen a positive impact on the factuality metrics when optimized for any of them. On the other hand, in most cases saw no/negative effects when summarization models are optimized for lexical overlap metrics such as ROUGE score. We hope that this work would promote further research into understanding interactions between the fact-aware methods and metrics to improve the quality of abstractive summarization.

6 Limitations

One of the primary limitations of this work is that fact-aware training uses hard thresholds to discard entire training examples, whereas there are abstractive summarization techniques that only modify the hallucinated content in the text while keeping the other portions intact. Additionally, although this is a part of our future work, another contribution could have been a novel metric that can ensemble characteristics of all the factuality metrics. Finally, in comparison to the baseline summarization models, fact-aware summarization models have larger inference times because of the pre-processing and post-processing required to choose relevant candidates either for training or as output.

References

Mario Barrantes, Benedikt Herudek, and Richard Wang. 2020. Adversarial NLI for factual correctness in text summarisation models. *CoRR*, abs/2005.11739.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *thirty-second AAAI conference on artificial intelligence*.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. *arXiv preprint arXiv:2104.09061*.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32.

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.

Alexander Richard 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.

Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084.

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.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. Tweetsumm-a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260.

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.

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.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51(3):371–402.

Daniel Kang and Tatsunori B Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731.

Wojciech Kryściński, 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.

Wojciech Kryściński, 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.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *arXiv preprint arXiv:2111.09525*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086.

Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained abstractive summarization: Preserving factual consistency with constrained generation. *arXiv preprint arXiv:2010.12723*.

Mani Maybury. 1999. *Advances in automatic text summarization*. MIT press.

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.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Feng Nan, Cicero dos Santos, Henghui Zhu, Patrick Ng, Kathleen Mckeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. 2021. 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.

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.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Leonardo FR Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. Factgraph: Evaluating factuality in summarization with semantic graph representations. *arXiv preprint arXiv:2204.06508*.

Kaiqiang Song, Logan Lebanoff, Qipeng Guo, Xipeng Qiu, Xiangyang Xue, Chen Li, Dong Yu, and Fei Liu. 2020. Joint parsing and generation for abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8894–8901.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

David Wan and Mohit Bansal. 2022. Factpegasus: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *NAACL 2022*.

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.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110.

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.

Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312*.

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.