Questioning the Validity of Summarization Datasets and Improving Their Factual Consistency

Yanzhu Guo[1], **Chloé Clavel**[2], **Moussa Kamal Eddine**[1], **Michalis Vazirgiannis**[1]
[1]LIX, École Polytechnique, Institut Polytechnique de Paris, France
[2]LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
{yanzhu.guo, moussa.kamal-eddine}@polytechnique.edu
chloe.clavel@telecom-paris.fr, mvazirg@lix.polytechnique.fr

Abstract

The topic of summarization evaluation has recently attracted a surge of attention due to the rapid development of abstractive summarization systems. However, the formulation of the task is rather ambiguous, neither the linguistic nor the natural language processing community has succeeded in giving a mutually agreed-upon definition. Due to this lack of well-defined formulation, a large number of popular abstractive summarization datasets are constructed in a manner that neither guarantees validity nor meets one of the most essential criteria of summarization: factual consistency. In this paper, we address this issue by combining state-of-the-art factual consistency models to identify the problematic instances present in popular summarization datasets. We release SummFC, a filtered summarization dataset with improved factual consistency, and demonstrate that models trained on this dataset achieve improved performance in nearly all quality aspects. We argue that our dataset should become a valid benchmark for developing and evaluating summarization systems.

1 Introduction

While the revolutionary success of the Transformer (Vaswani et al., 2017) architecture has drawn a surge of attention to automatic summarization, most research has been focused on improving performance metrics of summarization models on a set of popular datasets. It is often taken for granted that these datasets provide representative examples of high quality summaries, and that models capable of producing summaries that are similar to the ones of the dataset are superior in the task of automatic summarization. *But is this really the case?*

Numerous works (Fabbri et al., 2021; Huang et al., 2020) have already pointed out deficiencies in the most widely employed summarization datasets (e.g. CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016) and XSUM (Narayan et al., 2018)). These datasets are typically composed of news articles automatically extracted from news websites, paired together with a highlight or introduction sentence which serves as the summary. However, the qualities we seek in highlights and introductions are fundamentally different from the ones we seek in summaries. It is thus time for the NLP community to take a step back, revisit the formulation of the summarization task and reconsider the appropriateness of currently employed datasets.

The goal of automatic summarization is to take an information source, extract content from it and present the most important content to the user in a concise form and in a manner sensitive to the user's or application's needs (Mani, 2001). In the scope of this paper, we solely focus on generic summarization (Nenkova and McKeown, 2011), where we make few assumptions about the audience or the goal for generating the summary. While the importance of the selected content and the conciseness of its form are rather subjective, there is one criterion for summarization that is certain: the summary content should be extracted from the information source. In other words, the summary should be factually consistent with the source document. This is however not the case with summaries constructed from highlights or introductions. It is common for highlights and introductions to contain information that are not mentioned in the main article or even exaggerate certain facts to achieve the goal of click baiting. Therefore, the models trained on these datasets perform the task of "pitch generation" rather than the intended summary generation. An example of

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5716 - 5727

such a reference summary is given in Table 1.

In this paper, we aim to identify these erroneous data samples by scoring them with factual consistency models. We push towards answering two main **research questions** :

Q1 *What kind of factual inconsistency can different factuality models capture and how can they be combined for better detection of erroneous samples?* This question is answered in Sections 3.2 and 5.1 .

Q2 *Does having more reliable datasets with more factually consistent reference summaries lead to better performing summarization models? And do the results only improve in factual consistency or also in other quality aspects such as informativeness and saliency?* We address this question in Section 5.4.

Our research aims to answer these questions by filtering out samples with problematic reference summaries. We achieve this by leveraging state-of-the-art factual consistency models. We test the performance of different models on different categories of factuality errors (Pagnoni et al., 2021) and eventually combine them for an optimized filtration methodology.

Our **contributions** are summarized below:

1. We prove the effectiveness of three state-of-the-art factual consistency models in detecting factuality errors. We use these models to affirm the factual consistency issue present in three of the most popular summmarization datasets. We devise a filtration methodology that combines different state-of-the-art factuality models and achieves better detection of misleading reference summary samples.

2. We release SummFC[1], a **Summ**arization dataset with improved **F**actual **C**onsistency. We prove that fine-tuning summarization models on this dataset leads to better performance on not only factuality but also other quality aspects. We believe that regularly amending issues in widely employed benchmark datasets should become a common practice in NLP.

[1]The dataset is publicly available at https://github.com/YanzhuGuo/SummFC.

Source Document: *Archibald, 22, dominated the event, with Dutch rider Kirsten Wild second and Belgium's Lotte Kopecky third. That came 24 hours after she won her third consecutive women's individual pursuit title, having gained silver in Thursday's elimination race. The Scottish cyclist won gold in the pursuit quartet at the Olympics in Rio.*
Factually Inconsistent Reference Summary: *british olympic champion katie archibald won omnium gold at the european track championships, her second title in two nights in paris.*

Table 1: Example of a reference summary from the XSUM dataset. The words marked out in red represent factually inconsistent information. The most common type of error for the XSUM dataset is content verifiability errors.

2 Related Work

Summarization Benchmark Datasets Summarization benchmark datasets are typically composed of a large number of news documents paired together with "gold-standard" human reference summaries. Unfortunately, the human reference summaries in these datasets are often constructed in a suboptimal way. Gehrmann et al. (2022) analyze a sample of 20 papers proposing summarization approaches published in 2021; they find 27 datasets that models were being evaluated on. The most popular ones, CNN/DM (Nallapati et al., 2016) and XSum (Narayan et al., 2018), were used five and four times respectively. However, both of these datasets have multiple issues. Taking the CNN/DM dataset for example, the construction is done by pairing an article with the bullet points written for it on the CNN and Daily Mail websites. This design works well for its initial use as a Question Answering dataset (Hermann et al., 2015), but does not function at the same level after its adaptation for summarization. Regarding the XSUM dataset, the main issue concerns its factuality. The reference summaries were never meant to be a real summary, there is thus no requirement for it to be faithful to the source article. An analysis of XSum finds that over 70% of reference summaries contain factual inconsistencies (Maynez et al., 2020).

Factual Consistency Models Thanks to the development of reference-free factual consistency metrics for summarization (Huang et al., 2021), the evaluation of factual consistency can now be performed automatically on a large scale. Since such metrics are generally model-based, we refer to them as factual consistency models, in order to distinguish them from the evaluation metrics in Section 5.3. Popular factual consistency models can generally be categorized into two different paradigms: Question Answering (Durmus et al., 2020; Wang et al., 2020) and Entailment Classification (Kryscinski et al., 2020; Goyal and Durrett, 2020). Question Answering models are based on the intuition that if we ask the same question to a summary and its source article, they should provide similar answers if the summary is factual. Entailment classification models rely on the idea that a factually consistent summary should be semantically entailed by the source article. They are usually trained models fine-tuned on either synthetic or human-annotated datasets. According to the FRANK benchmark (Pagnoni et al., 2021), Entailment classification models perform significantly better than the Question Answering ones. In addition, some text generation evaluation metrics (Zhang et al., 2019; Yuan et al., 2021) falling out of these two paradigms are also proven to perform well as factual consistency models (Pagnoni et al., 2021).

Improving Factual Consistency The idea defended in this paper is that factual consistency models allow us to discover issues in current datasets and eventually release improved versions. Dataset quality in summarization has only started receiving attention recently and only a few papers have attempted to make efforts in this direction. Gehrmann et al. (2021) release an improved version of XSUM that filters the dataset with a BERT-based classifier fine-tuned on 500 document-summary pairs, manually annotated for faithfulness (Maynez et al., 2020). However, the classifier is fairly naive with only a single classification layer added after the BERT model and they did not compare the performance of models trained respectively on the original and improved datasets. Along similar lines, Matsumaru et al. (2020) build a binary classifier for detecting untruthful article-headline pairs and filter a headline generation dataset. Nan et al. (2021) also perform filtration for summarization datasets but both the filtration and evaluation methodologies are limited to the entity level. Goyal and Durrett (2021) identify non factual tokens in the XSUM training data with the Dependency Arc Entailment (DAE) model. They mask these tokens corresponding to unsupported facts and ignore them during the training step. While this approach does allow for improved summarization quality, we see potential issues arising when isolated ignored tokens are confronted with the contextual nature of the Transformer architecture. Filtering out entire samples with detected inconsistency is thus the preferred solution.

We present the first work on a systematic investigation of an optimal methodology for improving factual consistency in summarization datasets. We are also the first to comprehensively analyze the effect of training data quality on summarization system outputs. We furthermore extend the range of analysis compared to the previous papers, also performing experiments on the recently released XL-SUM dataset (Hasan et al., 2021).

3 Analyzing Factual Consistency Models

In this section, we present the factual consistency models we employ for detecting factual errors in the reference summaries. We introduce three state-of-the-art models that each rely on a distinctive mechanism. We rely on human annotations from the FRANK benchmark (Pagnoni et al., 2021) to validate the effectiveness of these models and analyze the variations in their performance when shifting dataset domain or error category. The FRANK benchmark is the most recent benchmark proposed for evaluating factual consistency models, it is the only benchmark to date that provides a categorization of factual errors.

3.1 Factual Consistency Models

In order to be comprehensive in our choice of factual consistency models, we run experiments on the top performing model from the FRANK benchmark as well as two other recently proposed models not included in the

FRANK benchmark. Each of the three models we choose depends on a different mechanism, thus they are expected to complement each other in detecting different types of errors.

BERTScore_Art Zhang et al. (2019) introduce BERTScore as an automatic evaluation metric for text generation. It computes a similarity score for each token in the candidate summary with each token in the reference summary leveraging BERT embeddings. It obtains the final scores by performing a greedy matching between tokens from both texts. Here we employ a slightly modified version BERTScore_Art, which directly compares a summary to the source article instead of the reference summary, for the sake of modeling factuality. We use the precision score which matches each token in the summary to a token in the article, instead of the recall score which does the opposite. This is the factuality model that obtained the highest correlation with human annotations in the FRANK benchmark (Pagnoni et al., 2021).

BARTScore Yuan et al. (2021) formulate the evaluation of generated text as a text generation task from pre-trained language models. The basic idea is that a high-quality summary should be easily generated based on the source article. The factuality score of a summary is calculated as its generation probability conditioned on the source document. The idea is operationalized using the encoder-decoder based model BART and thus named BARTScore. It builds a connection between the pre-training objectives and evaluation of text generation models. BARTScore is shown to produce state-of-the-art results for factuality evaluation on the SummEval Dataset (Fabbri et al., 2021).

DAE (Dependency Arc Entailment) Goyal and Durrett (2020) propose to decompose summaries into dependency arcs and train an entailment model that makes independent factuality judgments for each dependency arc of the summary. The judgements made by the model are binary and we use the probability for the factual class in order to obtain a continuous score. The arc-level judgements are then aggregated into summary-level by computing the mean score for all arcs present in the summary. The initial work proposes to train the entailment model on synthetic datasets. However, they later extend their method by training it on human annotations and achieve improved performance (Goyal and Durrett, 2021). This extended version of DAE is proven to be the best performing entailment-based model tested on the factuality dataset introduced in Falke et al. (2019).

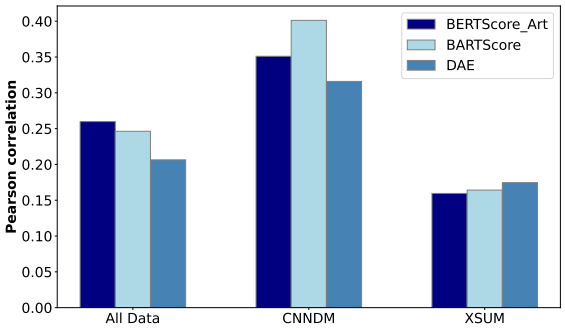


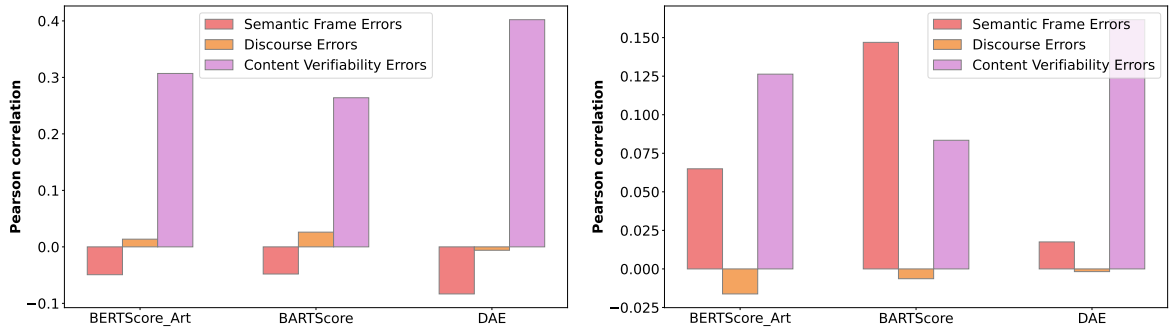
Figure 1: Partial Pearson correlation on different datasets. The three models demonstrate similar performance. All models have higher correlation on the CNN/DM dataset.

3.2 Model Validation Using the FRANK Benchmark

The FRANK benchmark proposes a linguistically motivated typology of factual errors for fine-grained analysis of factuality in summarization systems: semantic frame errors, discourse errors and content verifiability errors. Semantic frames refer to the schematic representation of an event, relation, or state and a *semantic frame error* is an error that only involves participants of a single frame. *Discourse errors* extend beyond single semantic frames and consist of erroneous relations between different discourse segments. *Content verifiability errors* arise when information in the summary cannot be verified against the source document.

The FRANK benchmark consists of human-annotated categorical error scores for 2250 model summaries. We use these annotations to compute partial correlation scores with human judgements for different models on different datasets and error types. We now examine these correlations.

Partial Pearson correlation on different datasets. Figure 1 shows the partial pearson correlation between different models and hu-



(a) XSUM (b) CNN/DM

Figure 2: Negative difference in partial Pearson correlation when flipping labels of an error type. Higher value indicates higher influence of the given error type in the overall correlation.

man judgments on document-summary pairs from different datasets. The FRANK benchmark includes 1250 pairs from the CNN/DM dataset and 1000 from the XSUM dataset. We observe that the three models achieve comparable performance across all datasets with BARTScore obtaining the best performance on CNN/DM, DAE obtaining the best performance on XSUM and BERTScore_Art on both of them combined.

Partial Pearson correlation on different error types. Figure 2 shows the variation in partial Pearson correlation when flipping the labels of a specific error type. A higher positive bar indicates that the given type of error has high influence in the overall correlation. In other words, the factual consistency model performs well in detecting an error type if it has a high positive bar. We observe that DAE performs the best for content verifiability errors on both of the datasets, but gives unsatisfactory results for the other two error types. This is expected as DAE scores each dependancy arc independantly and thus cannot detect erros spanning across different arcs. BERTScore_Art and BartScore are both able to achieve positive results for discourse errors on XSUM, and for semantic frame errors on CNN/DM. However, it is worth pointing out that all of the three models generally perform the worst for discourse errors, indicating that this is a challenging future direction for work on factuality models.

Due to the relative strength of each model in identifying different types of errors on different datasets, we choose to use all three of them in analyzing the summarization datasets in Section 4. We further combine these three models in our filtration process in Section 5.1.

4 Examining Summarization Benchmark Datasets

To question the validity of current summarization benchmark datasets, we first provide an overview of the context in which each dataset was created and the methodology employed in the construction procedures. In addition, we make use of the previously mentioned factual consistency models to evaluate the factuality of human reference summaries collected in each dataset. Our selection of summarization datasets is based on their popularity in recently published papers introducing new summarization systems. We believe that it is crucial to examine their validity as they play a fundamental role in both the development and evaluation of new systems. The most popular benchmark datasets according to our criteria are CNN/DM (Hermann et al., 2015) and XSUM (Narayan et al., 2018). We also include the very recently released dataset XL-Summ (Hasan et al., 2021). It is not yet widely employed but including it alongside the other two can help us inspect the most recent advances made in the field of summarization dataset creation. *Does the most recent benchmark dataset exhibit any improvement on factuality in comparison to the previous ones with well-known issues?*

4.1 Summarization Benchmarks

We present the three summarization benchmark datasets in the order of their time of release. We observe an evolution in the con-

struction methodology over time.

CNN/DM The CNN/DM dataset (Hermann et al., 2015) was initially constructed as a Question Answering dataset composed of newswire articles in English and their corresponding highlights from the two platforms CNN and Daily Mail. Cheng and Lapata (2016) later converted it into a summarization dataset by simply concatenating these highlights into summaries. It has now become the most broadly employed summarization dataset for the English language. However, the summaries formed from the concatenation of bullet points exhibit low degrees of abstraction and coherence, which are both highly desirable qualities for abstractive summarization systems. The dataset consists of 311,971 document-summary pairs.

XSUM Narayan et al. (2018) create another large-scale dataset for abstractive summarization by crawling online articles from the BBC platform. They take the first line of an article as the summary and the rest of the article as the source document. This method for annotating summaries guarantees high levels of abstraction but creates other issues as the first line of an article is often not written to be the summary. It might either include meta-information such as the author and publication date or serve as background introduction thus containing information never mentioned again in the rest of the article. XSUM contains 226,711 document-summary pairs.

XL-Sum XL-Sum (Hasan et al., 2021) is a recently introduced abstractive summarization dataset containing document-summary pairs also extracted from the BBC platform. The automatic annotation strategy of summaries is similar to that of XSUM. However, they find the first line of many articles to contain meta-information and thus annotate the bold paragraphs instead as the summaries. The dataset covers 44 languages, for many of which no public summarization datasets were previously available. Summaries contained in XL-Sum are also highly abstractive. The English subset of this dataset contains 329,592 document-summary pairs.

4.2 Factual Consistency of Summarization Benchmarks

We focus on evaluating the factual consistency of human reference summaries contained in each of the summarization benchmark datasets. We often assume that human references are "gold standards" but there is a lot to question in their validity considering how they are annotated.

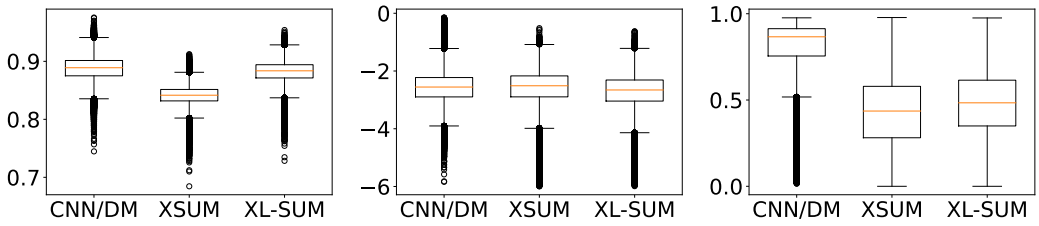
Here, we perform an analysis on the human references using the three factuality models introduced in Section 3.1: BERTScore_Art, BARTScore and DAE. The results are shown in Figure 3. We remark that factuality scores produced by all three models rank CNN/DM as the most factual dataset by a large margin. This is due to its low level of abstraction and thus do not indicate its superiority as an abstractive summarization dataset. However, it is also worth pointing out that the improvement achieved betweeen XSUM and the recently created XL-SUM is not significant. By qualitatively analyzing the lowest ranking summaries scored by each model from each dataset, one can confirm their non factuality. A representative non factual example for XSUM is shown in Table 1. Examples for the two other datasets are shown in Appendix Section A.

5 Introducing the SummFC Dataset

The approach we choose to address the factuality problem in these summarization datasets is to filter the samples by their factual consistency scores obtained in Section 4.2. We apply this idea to construct **SummFC**, the new **Summ**arization benchmark dataset with improved **F**actual **C**onsistency. In the remainder of this section, we first present our filtration methodology, followed by statistics of the SummFC dataset. We then introduce the evaluation metrics used in the comparison of summaries generated by models trained on the original benchmarks and models trained on SummFC. Finally, we perform experiments on the three datasets introduced in Section 4.1 and observe advantages of the SummFC dataset.

5.1 Filtration Methodology

In Figure 4, we illustrate our pipeline for filtration and evaluation of the three datasets. The



(a) BERTScore (b) BARTScore (c) DAE

Figure 3: Box plot of factuality scores obtained for the three datasets.

		#Samples	Selection Ratio	Document Length	Summary Length
CNN/DM	Full train set	287,113	*	788	54
	SummFC selection	153,666	53.52%	726	56
XSUM	Full train set	204,045	*	430	23
	SummFC selection	118,427	58.03%	405	23
XLSUM	Full train set	306,522	*	530	24
	SummFC selection	161,200	52.59%	491	24

Table 2: Statistics of the SummFC dataset.

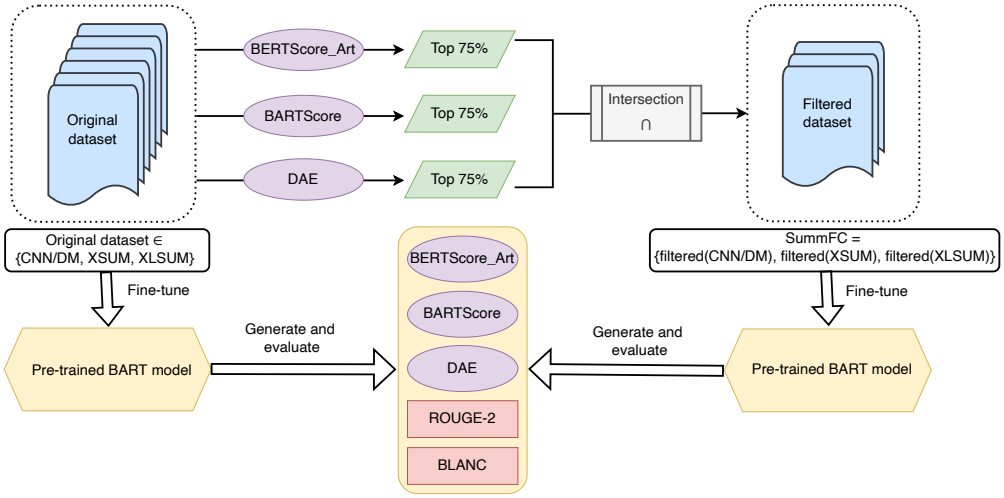


Figure 4: Pipeline for creating and evaluating the SummFC dataset.

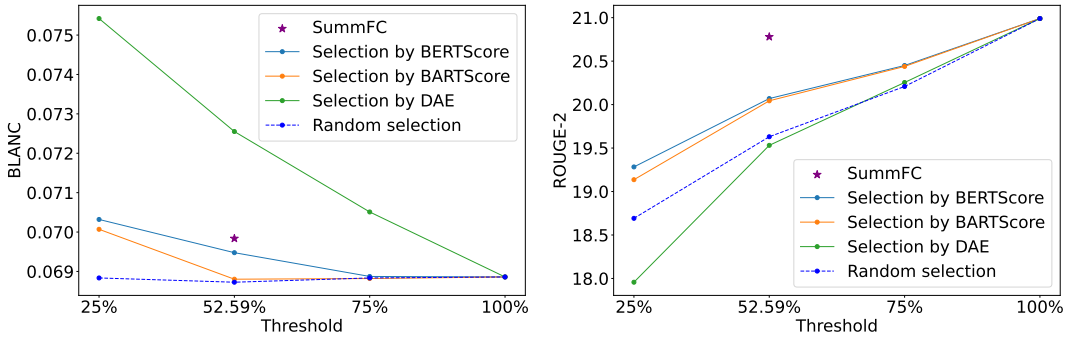
filtration methodology is based on the factuality models BERTScore_Art, BARTScore and DAE. For each of the three datasets, we filter out the bottom 25% of document-summary pairs scored by each of the three models. In other words, we only keep the intersection between the top 75% of samples scored by each factuality model. This choice is made because the three factuality models are shown to complement each other in the detection of erroneous samples (see Section 3.2). Experiments in Section 5.4 show that datasets filtered by combined models outperform single model filtration with the same threshold.

5.2 The SummFC Dataset

In Table 2, we show statistics for the SummFC datasets. For all three datasets, the selection ratio of samples to retain is between 50% and 60%. In the context of our filtration methodology removing the lowest scored 25% of samples for each factuality model, this final selection ratio proves that there is a high degree of overlap between the samples different factuality models choose to filter out. We also report the average length of documents and summaries in each dataset. We compute length as the number of words (instead of tokens) included in the text. We might expect to achieve higher factual consistency scores for shorter summaries, as they

		BERTScore_Art	BARTScore	DAE	BLANC	ROUGE-2
CNN/DM	Random 53.52%	0.9332	-2.662	0.9046	0.1538	22.80
	Full train set	0.9346	**-2.577**	0.9001	**0.1593**	**23.77**
	SummFC selection	**0.9364**	-2.600	**0.9073**	0.1556	23.67
XSUM	Random 58.03%	0.8973	-2.386	0.4134	0.07288	20.27
	Full train set	0.8944	-2.452	0.4201	0.07316	**21.78**
	SummFC selection	**0.8982**	**-2.383**	**0.4473**	**0.07378**	21.50
XLSUM	Random 52.59%	0.8979	-2.537	0.4612	0.06872	19.63
	Full train set	0.8961	-2.594	0.4695	0.06886	**20.99**
	SummFC selection	**0.8989**	**-2.501**	**0.5120**	**0.06983**	20.78

Table 3: Results for BART models fine-tuned on different selections of datasets. The statistical significance ($p < 0.05$) of all results are confirmed using the Wilcoxon signed-rank test.



(a) BLANC (b) ROUGE-2

Figure 5: Results for BART models fine-tuned on different subsets of the XLSUM dataset, filtered by different factuality models with varying thresholds.

naturally contain less information. However, we show that our filtration methodology does not particularly favor samples with shorter reference summaries.

5.3 Evaluation Metrics

To compare the quality of summaries produced by models fine-tuned on the the original datasets and the SummFC selection, we employ the following evaluation metrics capturing different quality aspects.

Factual Consistency Models For testing the improvement in **factual consistency** of the produced summaries, we employ the same factuality models as we use for filtration in Section 3.1: BERTScore_Art, BARTScore and DAE.

BLANC BLANC (Vasilyev et al., 2020) is a reference-free evaluation metric for summarization. It is defined as a measure of how well a summary helps an independent, pre-trained language model while it performs its language understanding task (masked token task) on a document. In other words, BLANC measures the **informativeness** of a summary in the context of document understanding. Here, we use the BLANC-help version, which uses the summary by concatenating it to each document sentence during inference.

ROUGE-2 ROUGE-2 (Lin, 2004) is a reference-based text generation metric computing the bigram overlap between reference and candidate summaries. While reference-free evaluation metrics are advantageous due to the issues in reference summaries, there are still quality aspects it cannot cover. There is currently no reference-free metric that is able to measure a summary's **saliency** (*i.e.* coverage of salient information). Thus we still need to compare the generated summaries to the reference ones with ROUGE-2, assuming that the reference summaries exhibit a high degree of saliency. We also process the samples in the test set with our filtration methodology when using ROUGE-2. For all the other evaluation metrics, we do not perform filtration for the

test set because they are reference-free and thus not influenced by misleading reference summaries.

5.4 Experiments and Results

We fine-tune the pre-trained BART-base model (Lewis et al., 2020) from the Transformers library (Wolf et al., 2020) on the different summarization datasets. Since our goal is not to advance summarization systems but to compare different benchmark datasets, we do not experiment with hyperparameter tuning. For all parameters except batch size, we use the default settings in the Transformers library. Both the training and prediction are done on a NVIDIA TITAN V GPU with a batch size of 8. We present results for models fine-tuned on the original benchmarks and SummFC in Table 3. We also create a random selection baseline for each dataset by uniformly sampling a random subset of samples with equal size to the SummFC selection.

We observe that models trained on the SummFC selection of XSUM and XLSUM achieve the highest scores for all reference-free evaluation metrics. Although the selection procedure is only based on factual consistency, we see that the summarization systems have also improved in the informativeness aspect as measured by BLANC. For the reference-based metric ROUGE-2, models trained on SummFC obtain comparable scores to those trained on the original dataset and beat the random baseline by a large margin. It is also interesting to remark that for some of the factuality scores, even the random baseline achieves better performance than the full train set. This further confirms the conclusion that too much erroneous training data hinders factual consistency in summarization models. The results on CNN/DM are consistent with the other two datasets, except for BARTScore and BLANC. We believe that our filtration methodology is slightly less effective on CNN/DM due to the fact that this dataset manifests the lowest degree of factual inconsistency. Another important reason is that the BART model used in BARTScore is also fine-tuned on CNN/DM, which makes it biased on this dataset. However, it is worth noting that SummFC is considerably smaller in size compared to the original ones. This means that *using SummFC, we can achieve better results on nearly all quality aspects while reducing training time and lowering the need for computational resources*.

Figure 5 shows metric scores obtained when fine-tuning on filtered XLSUM datasets with single factuality models at different thresholds. We show scores for BLANC and ROUGE-2 because these two metrics were not used during filtration. We also create a baseline by randomly selecting samples with proportions equal to the tested thresholds. As the filtration criteria becomes stricter, BLANC scores increase while ROUGE-2 scores decrease. Our final filtration threshold is chosen as a compromise between the optimization of these two scores. We also observe that there is no single factuality model that performs the best for both scores. *In general, the SummFC combined filtration strategy outperforms single factuality models at the same threshold.*

6 Conclusion

In this paper, we demonstrate that popular summarization datasets suffer from the lack of factual consistency and that summarization models trained on these datasets are not adequate for the task of abstractive summarization. We show that this problem can be solved by filtering the benchmark datasets with scores from factual consistency models. We propose a filtration methodology combining three state-of-the-art factual consistency models and introduce the SummFC dataset. SummFC is a unified **Summ**arization benchmark consisting of **F**actually **C**onsistent samples chosen from CNN/DM, XSUM and XLSUM. Experiments indicate that, in general, models trained on the smaller SummFC generate summaries with higher quality than models trained on the larger original datasets. Our findings suggest that more deliberate considerations should be made in the construction of benchmark datasets and that continuous revisions for the already existing ones are particularly necessary. However, constructing adequate benchmark datasets with textual contents and labels matching the initial formulation of NLP tasks remains an open question, major obstacle, and unresolved issue for the whole community.

Limitations

In this work, we restrict ourselves to the most popular summarization datasets. The three analyzed datasets share many similarities with each other and thus cannot account for diversity. They are all single-document single-reference English language summarization datasets in the news domain. Due to the limited sequence length accepted by Transformer-based factual consistency models, our filtration methodology cannot be generalized for longer document-summary pairs which are frequently present in other domains such as scientific articles and creative writing. The methodology also cannot be generalized for other languages. Factual consistency models such as DAE are trained on annotated datasets which only exist for the English language. This provides motivation for improving the generalization of factual consistency models.

Another limitation of our work is the lack of human evaluation. We prove the superiority of the SummFC dataset by evaluating the generated summaries with automatic metrics. Although these metrics achieve state-of-the-art correlations with human annotated scores, they are still known to significantly differ from human judgements. While the factual consistency models that we employ also represent the current state-of-the art, it is far from guaranteed that they are able to identify all the factuality errors. It is only with human evaluation that we can provide a complete picture of where SummFC falls on the full dataset quality continuum. There is undoubtedly still more work to be done in continuing to refine datasets to actually measure summarization.

Ethics Statement

We believe that research on improving the factual consistency of summarization systems can create positive social impact. We live in an era of information explosion and everyone, to some degree, relies on summarization to process the information overload. It is our responsibility to guarantee the greater public's access to truthful information.

Our research also brings positive impact on the environmental aspect. Training on smaller and higher quality datasets significantly reduces the consumption of computational energy while also boosting performance.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback and insightful comments. The first and last authors were supported by the ANR HELAS chair (ANR-19-CHIA-0020-01). The second author was supported by the Télécom Paris research chair on Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS).

References

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

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

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

Tobias Falke, Leonardo F. R. 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, Florence, Italy. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly

Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

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.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

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.

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.

Inderjeet Mani. 2001. *Automatic summarization*, volume 3. John Benjamins Publishing.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346, Online. 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.

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, Berlin, Germany. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. 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.

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.

Ani Nenkova and Kathleen McKeown. 2011. *Automatic summarization*. Now Publishers Inc.

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.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Examples of Factually Inconsistent Reference Summaries

Source Document:
By . Leah Simpson . PUBLISHED: . 16:46 EST, 19 July 2012 . | . UPDATED: . 02:31 EST, 20 July 2012 . With the season finale airing on Sunday night, The Bachelorette star Emily Maynard is already making arrangements to extend her 15 minutes of fame - with a move to Hollywood on the cards. ..... But she's really excited to get the date, location and all of the details set so that she can marry Jef and having it air on TV fits in perfectly with her plans.' Happy couple: Emily is engaged to get married to Jef Holm apparently .
Factually Inconsistent Reference Summary:
Bachelorette spoiler alert .

Table 4: Example of a reference summary from the CNN/DM dataset. It is more of a click-bait title for attracting attention than a real summary.

Source Document:
Bosses said the move was part of an efficiency drive, with 10 posts set to go in Haverfordwest and 20 at its plant in Aspatria, Cumbria. "We recognise that the impact of these proposed changes is significant for the people affected and we are committed to treating people with respect and consideration," said a spokesman. A staff consultation starts this week.
Factually Inconsistent Reference Summary:
A total of 30 jobs are under threat at two First Milk creameries in Pembrokeshire and the Lake District.

Table 5: Example of a reference summary from the XL-SUM dataset. Information in summary not explained in the source document.