Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-off in Abstractive Summarization

Anonymous ACL submission

Abstract

Despite recent progress in abstractive summarization, systems still suffer from faithfulness errors. While prior work has proposed models that improve faithfulness, it is unclear whether the improvement comes from an increased level of extractiveness of the model outputs as one naive way to improve faithfulness is to make summarization models more extractive. In this work, we present a framework for evaluating the *effective faithfulness* of summarization systems, by generating a *faithfulness-abstractiveness trade-off curve* that serves as a *control* at different operating points on the abstractiveness spectrum. We then show that the Maximum Likelihood Estimation (MLE) baseline as well as recently proposed methods for improving faithfulness, fail to consistently improve over the *control* at the same level of abstractiveness. Finally, we learn a selector to identify the most faithful and abstractive summary for a given document, and show that this system can attain higher faithfulness scores in human evaluations while being more abstractive than the baseline system on two datasets. Moreover, we show that our system is able to achieve a better faithfulness-abstractiveness trade-off than the *control* at the same level of abstractiveness.

1 Introduction

Generating abstractive summaries of documents has been a long-standing goal of summarization. While there has been tremendous progress towards this goal (Kryściński et al., 2018; Dong et al., 2019; Zhang et al., 2019; Lewis et al., 2020), abstractive summarization systems still suffer from faithfulness errors, hallucinating information that is not present in the original text. This has led to an increased research in faithfulness evaluation of summarization systems (Falke et al., 2019; Kryscinski et al., 2020; Durmus et al., 2020) as well as methods to improve faithfulness of generated summaries (Kang and Hashimoto, 2020; Chen et al., 2021). Intuitively, one straightforward way of improving faithfulness of generated summaries is to copy a larger amount of content from the source article (i.e. more extraction). Thus, any methods that increase the level of extractiveness, whether intentionally or not, would improve faithfulness. Without reported extractiveness, it is unclear whether prior improvements mainly arise from increased extractiveness. We argue that in order to make progress in abstractive summarization, it is important to tease apart faithfulness improvements due to increased extractiveness versus improvements due to improved abstraction.

In order to tease this apart, we develop a framework for evaluating progress in faithfulness, by considering the *effective faithfulness*, i.e. the improvement in faithfulness over a baseline system (*control*) operating at the same level of extractiveness. In particular, we split the training examples into different groups by the extractiveness of the summary, and train the *control models* on each group. Each of these models corresponds to a specific tradeoff between abstractiveness and faithfulness, forming a *trade-off curve* indicating how much faithfulness can be improved solely by increasing extractiveness. Systems that improve *effective faithfulness* should lie above this curve.

Using this framework, we show that the improved faithfulness of recently proposed methods comes mainly from an increased extractiveness. We then conduct further analysis to explore whether it is possible to have a system that can be both more abstractive and more faithful than the MLE baseline system. We train a selector on a small set of human-annotated data that, given a set of output summaries with varying levels of extractiveness, picks the most abstractive output that is faithful to the source. Our proposed system is both more abstractive and faithful than the MLE baseline. Moreover, we show that our system is able

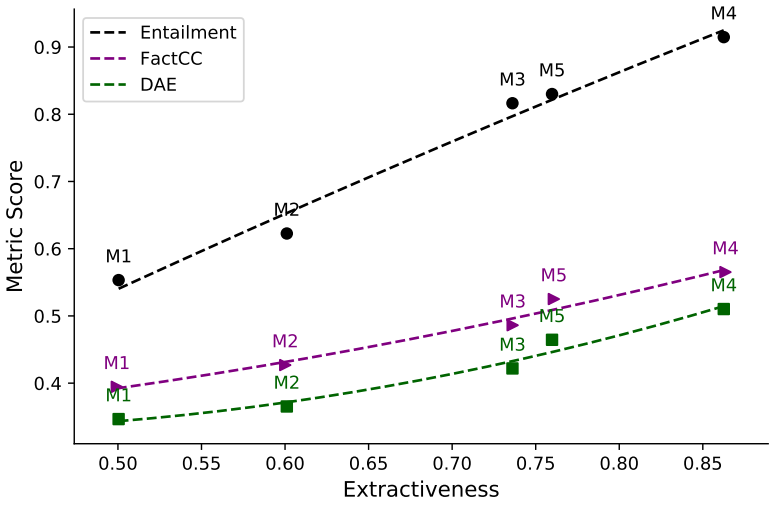


Figure 1: Extractiveness of generated outputs versus metric scores for Entailment, FactCC and DAE on the Gigaword dataset. We use *coverage* defined in Grusky et al. (2018) to measure extractiveness, where summaries with higher coverage are more extractive. We observe that automated metrics of faithfulness are positively correlated with extractiveness.

to improve the *effective faithfulness*, achieving a better trade-off than the *control* at the same point on the abstractiveness sepctrum.

To summarize, our contributions are as follows:

1. We present a framework to evaluate the progress in improving *effective faithfulness* of the models considering the *control* at the same level of extractiveness.

2. We illustrate the importance of considering *effective faithfulness* by showing that recently proposed methods for improving faithfulness are able to attain higher faithfulness scores than the MLE baseline, but do not consistently improve over the *control curve*, indicating that most of their improvements come from generating more extractive outputs, on average.

3. We propose a selector that picks the abstractive, faithful summary from a set of possible summaries, and show that this method gets higher effective faithfulness compared to the existing methods.

2 Dataset

We conduct our study on two datasets, one from the news domain, and one from a non-news domain. For the news domain dataset, we decided against using the popular CNN/Dailymail dataset since its reference summaries tend to be very extractive (Kedzie et al., 2018; Bommasani and Cardie, 2020), making it a poor choice for studying faithfulness in abstractive summarization. Similarly, we also decided against using XSum, another popular news summarization dataset, since almost 77% of the gold reference summaries contain hallucinations (Maynez et al., 2020). Instead, we opted for Gigaword and Wikihow, which are datasets with substantial abstraction without as much hallucination problems as XSum. Gigaword reference summaries have substantially less hallucinations than XSum (Kang and Hashimoto, 2020), and WikiHow summaries tend to be of a higher quality since they are written and curated by humans (Koupaee and Wang, 2018; Ladhak et al., 2020).

Wikihow (Koupaee and Wang, 2018) is a dataset of how-to articles covering a diverse set of topics, collected from the wikihow.com website. Each article contains several paragraphs detailing step by step instructions for a procedural task. There are about 12M such paragraphs in the dataset, paired with a one sentence summary.

Gigaword (Rush et al., 2015) is a headline generation dataset that contains around 4M examples, extracted from news articles that were collected as part of the Gigaword corpus (Graff et al., 2003). The model is tasked with generating the headline of the article given the first sentence.

2.1 Dataset Extractiveness

We follow the process detailed by Grusky et al. (2018), and use *extractive fragment coverage* and *extractive fragment density* as the measures of extractiveness of a given summary. Henceforth we will refer to these as coverage and density respectively. Coverage is the percentage of words in a summary that are from the source article. Density is the average length of the text spans copied from the document that are contained in the summary. A summary that copies larger chunks of text from the source article will have a higher density.

3 Analysis on Metrics of Faithfulness

Recent studies of faithfulness evaluation have proposed model-based automated metrics to detect whether a given summary is faithful to the source article. For example, Falke et al. (2019) have studied using pretrained entailment based methods to assess the probability of the generated output being entailed by the source article. Kryscinski et al. (2020) augment hallucinated summaries by applying rule-based transformations to the document sentences and train a BERT-based model to classify whether the generated output is faithful. Goyal and Durrett (2021) have collected fine-grained annotations to study word-, dependency- and sentence-level faithfulness and use these annotations to train a factuality detection model.

Figure 1 shows the relationship between the average coverage of the generated outputs (extractiveness) vs. average metric scores (faithfulness) assigned to various abstractive summarization models trained on Gigaword. We observe that there is a positive correlation between extractiveness and faithfulness scores, as models whose generated summaries have a higher average coverage tend to also get higher scores for each of the faithfulness metrics. This correlation between exractiveness and faithfulness makes it unclear whether a model gets higher factuality scores simply because it is more extractive or it is capable of generating faithful summaries at the original level of extractiveness. This highlights the need for accounting for extractiveness in order to compare faithfulness across different abstractive summarization systems.

4 Evaluating Effective Faithfulness

Given that extractiveness is confounded with faithfulness, we propose a framework for evaluating *effective faithfulness*, which takes into account the extractiveness of a system. In order to do this, we first need to determine the faithfulness of a system operating at a given level of extractiveness. We call this the *Faithfulness-Abstractiveness Tradeoff* and we describe it further in §4.1. The *effective faithfulness* of a system is then simply the relative difference between the faithfulness score assigned to the system, and the score of a system operating with the same average extractiveness according to the trade-off curve.

4.1 Faithfulness-Abstractiveness Tradeoff

In order to understand the effectiveness of a proposed system for improving faithfulness, we need to be able to account for its extractiveness. We finetune pre-trained BART models (Lewis et al., 2020) for different levels of extractiveness, without any explicit recourse for improving faithfulness. We then use these systems to create a *faithfulness-abstractiveness trade-off curve* that can serve as a control to measure the *effective faithfulness* of summarization systems. Models that improve *effective faithfulness* should lie above the *faithfulness-abstractiveness trade-off curve*

In particular, we sub-sample the training data into extractiveness quartiles by computing the coverage of the references with respect to the source articles. We then fine-tune the MLE baseline model on each of these quartiles to obtain models with varying level of extractiveness. We then collect human annotations for faithfulness of the summaries generated by each of these models as well as the MLE baseline for a random sample of 200 articles. We collected three annotations per example on Amazon Mechanical Turk asking whether an output is faithful or unfaithful with respect to the corresponding source article. We then compute the percentage of annotators that selects "faithful", and use this as the faithfulness score for each example.

Table 2 shows the coverage and faithfulness scores for the baseline and each of the models fine-tuned on the data quartiles, where Q1 is the *most abstractive* and Q4 is the *most extractive* quartile. We observe that the models that are fine-tuned on more extractive quartiles produces output with significantly higher coverage and faithfulness scores. The baseline model generates relatively extractive output with coverage closest to Q3 on both Gigaword and Wikihow. Furthermore, we observe that the baseline model has a higher coverage than the model fine-tuned on Q3 but it has lower faithful-

Article	Once you decide what to outsource, look for the right contractors. Start by asking for referrals from your own professional network. Talk to other business owners and professionals about how and where they outsource. You can also check professional associations or trade groups field in which you are trying to outsource work. Use other social media platforms such as Facebook or Twitter to advertise what you are looking for. Alternately, you can connect with contractors and freelancers on sites such as eLance, Guru and oDesk. These websites allow business owners to place an ad that describes what kind of work they need to have done, and contractors respond with their qualifications and rates. Send each potential provider the same bid document you prepared so that you can more easily compare their offers.
Baseline	Search for contractors and freelancers to outsource the work.
Q1	Conduct an initial search for qualified contractors and freelancers.
Q2	Search for qualified contractors and freelancers to work on your project.
Q3	Search for contractors and freelancers to do the work.
Q4	Look for contractors and freelancers to bid on the work.

Table 1: Example summaries generated by the baseline and quartile models for the article "*How to Outsource Small Business Tasks*" from Wikihow dataset. The tokens that do not appear in the source article are indicated by green.

Dataset	Model	Coverage	Faithfulness
Gigaword	Baseline	76.12	83.33
	Q1	50.25	71.83
	Q2	60.57	79.50
	Q3	73.64	86.67
	Q4	86.94	89.17
Wikihow	Baseline	88.28	82.52
	Q1	81.34	67.82
	Q2	85.34	76.21
	Q3	87.59	80.35
	Q4	90.19	91.08

Table 2: Coverage and faithfulness values of the MLE baseline and each quartile model for Gigaword and Wikihow. Quartile models with higher coverage have higher faithfulness scores.

Dataset		Cov.	Faithfulness
Gigaword	Baseline	76.12	83.33
	bf	77.74	89.57
	bfe	61.87	90.67
	qfe	63.55	98.00
Wikihow	Baseline	82.52	88.28
	bf	83.95	92.20
	bfe	70.52	91.32
	qfe	72.58	98.61

Table 3: Oracle coverage and faithfulness values for Gigaword and Wikihow. The oracle analysis suggests that being able to control for extractiveness can allow us to build systems that mitigate the trade-off.

ness score for Gigaword.

Table 1 shows an article from the Wikihow dataset and corresponding output summaries generated by the MLE baseline and each of the quartile models. We observe that the generated summaries are very similar in meaning; however, the output generated by the Q1 model includes a higher number of novel words (i.e. lower coverage) compared to the other models while staying faithful to the article. Conversely, Q4 model has a coverage of 1 in this example; all the words generated by this model are from the source article. On average, the Q1 model generates output that is more abstractive and less faithful while Q4 generates output that is more extractive and more faithful.

5 Mitigating the Trade-off

5.1 Oracle Experiments

We first aim to understand whether it is possible to mitigate the faithfulness-abstractiveness tradeoff by designing several oracle experiments where we have access to human judgments.

baseline + faithfulness (bf). We use the output from the MLE baseline model if it is faithful (i.e. at least two out of three annotators agree that the output is faithful). If the baseline output is not faithful, we select the output from the quartile model that is more extractive than the baseline to see whether we can have a similar coverage as the baseline but preserve faithfulness. **baseline + faithfulness-**

	Gigaword		Wikihow	
	Coverage	Faitfulness	Coverage	Faithfulness
Baseline	76.12	83.33	82.76	86.94
Loss Truncation	79.55	87.17	84.93	87.84
DAE	78.23	86.33	84.15	88.83
Selector-ROC (Ours)	64.58	84.17	78.67	87.84
Selector-F_β (Ours)				
β				
0.5	54.77	76.83	64.24	79.82
0.4	59.79	81.67	67.81	81.71
0.3	60.72	82.00	68.53	83.15
0.2	68.38	86.00	78.67	87.84
0.1	79.92	88.00	84.72	89.19

Table 4: Coverage and faithfulness scores for the baselines and our proposed methods. We show that with our method we are able to get models that are both more faithful and more abstractive than the baseline.

extractiveness (bfe). This oracle system behaves similar to the one described above when the MLE baseline output is unfaithful. However, rather than always selecting the baseline output when it is faithful, we pick the output from the quartile model that is more abstractive than the baseline whenever it is also faithful according to human judgement. **quartile + faithfulness-extractiveness (qfe).** Amongst the output of all four quartile models, we pick the most faithful output with the highest level of abstractiveness to understand whether it is possible to generate abstractive output while remaining faithful.

Analysis. Table 3 shows the coverage and faithfulness of the MLE baseline and each of these oracles for Gigaword and Wikihow. We observe that it is possible to be more faithful than the baseline at a similar level of abstractiveness (bf). Furthermore, we can be more abstractive than the baseline while being more faithful (bfe). Selecting the most faithful and abstractive output from the quartile models achieves a really high faithfulness score (≈98%) while having significantly less coverage than the baseline. This oracle analysis suggests that it should be possible to build models that can mitigate the faithfulness-abstractiveness trade-off by controlling the level of extractiveness. Given this, we further explore whether we can learn a selector that is capable of doing this selection automatically to mitigate the faithfulness-abstractiveness trade-off.

5.2 Loss Truncation

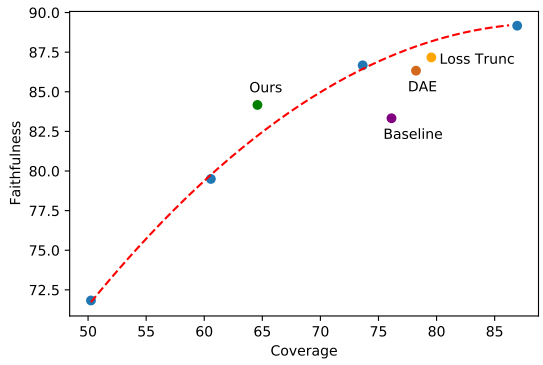
Kang and Hashimoto (2020) have proposed a method to adaptively remove high loss examples to optimize the distinguishability of samples from the model and the reference. They have shown that the samples generated by this Loss Truncation model achieves higher factuality ratings compared to the baseline methods. We study this method to understand where it lies in terms of faithfulness-abstractiveness trade-off and whether it can achieve a improved *effective faithfulness* over the *control*.

5.3 Dependency Arc Entailment (DAE)

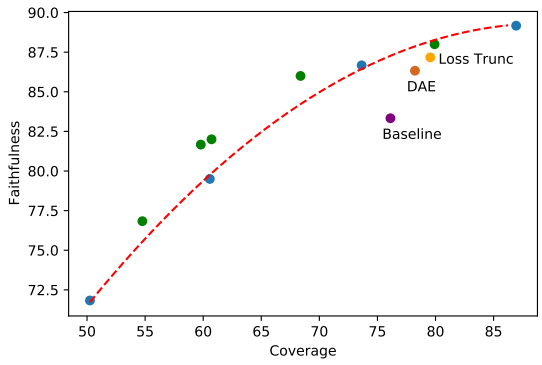
Goyal and Durrett (2020) have proposed a factuality evaluation metric (DAE) that evaluates whether each dependency arc in the generated output is consistent with the input. They show that their proposed metric works better than existing factuality metrics, while also being able to localize the parts of the generated output that are non-factual. Goyal and Durrett (2021) take advantage of DAE's ability to localize factuality errors, and train a summarization model only on the subset of tokens that is deemed factual according to the DAE metric. We follow their methodology to train summarization models, and assess them using our evaluation framework.

5.4 Selector Model

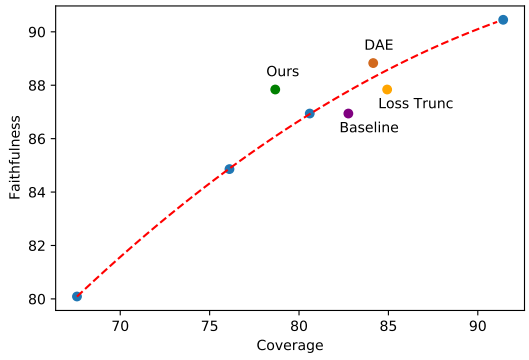
We aim to understand whether we can build a model that achieves a better *effective faithfulness* than Loss Truncation. We propose a selector that can identify the most abstractive but faithful output to improve this trade-off. We first generate four possible candidate summaries using the quartile models for each example in the validation set.



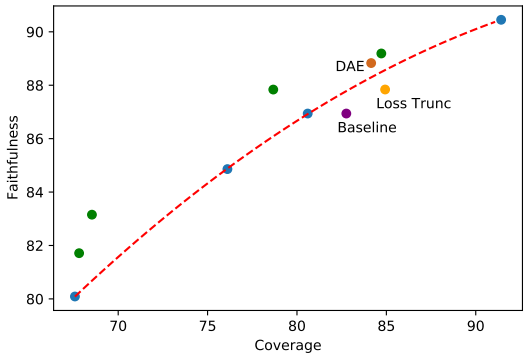
(a) Selector-ROC and the baseline trade-off on **Gigaword**.



(b) Selector-F$_\beta$ and the baseline trade-off on **Gigaword**.



(c) Selector-ROC and the baseline trade-off on **Wikihow**.



(d) Selector-F$_\beta$ and the baseline trade-off on **Wikihow**.

Figure 2: Faithfulness-Abstractiveness trade-off curves. The blue dots represent the quartile models used to generate the curve. The purple dot corresponds to the MLE baseline. DAE and Loss Truncation are depicted by the brown and orange dots respectively. The green dots correspond to our proposed systems.

This results in outputs with varying levels of extractiveness. We then use a selector system that first assigns faithfulness scores to each of these summaries and then selects the most abstractive, faithful output. We do 10-fold cross validation and fine-tune a FactCC model (Kryscinski et al., 2020) on the data we collected to generate the trade-off curve[1] and pick the faithfulness thresold that maximizes the area under the ROC curve (**Selector-ROC**). We then use this model to predict the faithfulness scores of the test folds. For each example, we select the most abstractive output that is considered faithful according to this model (if the faithfulness score is above the tuned thresold).

Instead of maximizing for the area under the ROC curve, we can also tune the faithfulness threshold to maximize F$_\beta$ scores (**Selector-F$_\beta$**). Using F$_\beta$ score with $\beta < 1$ allows us to assign a higher weight to the precision of our selector which would result in outputs with higher coverage and faithfulness.

We find that the fine-tuning step is important since pre-trained faithfulness models are trained on a different set of examples and do not transfer well to our datasets. This is consistent with the findings of Goyal and Durrett (2021).

5.5 Results

Table 4 shows the coverage and faithfulness results for the MLE baseline, Loss Truncation, DAE, and the selectors. We observe that as we use smaller values for β for Selector-F$_\beta$, we get more extractive and more faithful outputs. This allows us to have a trade-off between faithfulness and abstractiveness. Moreover, with both Selector-ROC and Selector-F$_\beta$, we produce output with less coverage but higher faithfulness scores than the MLE baseline. For Wikihow, Selector-ROC produces outputs with lower coverage but similar faithfulness scores to Loss Truncation. We can further obtain a higher faithfulness score at a similar coverage level as DAE and Loss truncation with Selector-F$_\beta$ with

[1]We collected annotation for 200 articles for each of the quartile models.

Article	If applicable, the description of any people who take part in your study should be extremely thorough. Each person should be identifiable within the research. Further, how people join and leave the study should be noted. If people were selected at random, or if they were family members, is important to the study. Be sure to consider various ethical concerns (e.g. risk and consent of participants) if people are involved in your research.
MLE Baseline	Describe who is involved in the study.
DAE	Identify the people who take part in the study.
Loss Truncation	Describe people who take part in your study.
Selector-ROC (Ours)	Describe all participants thoroughly and with care.
Article	Because diarrhea frequently causes dehydration, it is crucial that patients with IBD remain hydrated. Drink at least 8 glasses of water every day (or 64 oz). Foods that have a high water content (like watermelon) can also count toward this minimum. If you have a severe attack of diarrhea, you are likely to lose electrolytes. In these cases, you might need to consume beverages such as Pedialyte or Gatorade to help replenish them [TRUNCATED] ...
MLE Baseline	Drink plenty of water to stay hydrated.
Loss Truncation	Drink plenty of water.
DAE	Drink Drink plenty of water to stay hydrated.
Selector-ROC (Ours)	Drink Drink plenty of fluids to stay hydrated.

Table 5: Example summaries generated by the MLE baseline, Loss Truncation and the selector model.

$\beta = 0.1$. For Gigaword, Select-ROC produces output with significantly lower coverage than Loss Truncation and DAE. Selector-F$_\beta$ produces output with similar coverage to Loss Truncation with a higher faithfulness score ($\beta = 0.1$).

It is important to understand whether models improve faithfulness by simply being more extractive or if they are able to improve *effective faithfulness*. In order to understand this, we measure whether the models get improvement in faithfulness over the *control* operating at the same level of extractiveness. In Figure 2, we plot the faithfulness-abstractiveness curve with the faithfulness and abstractiveness of the quartile models. If a model lies above this curve, it improves the *effective faithfulness*. If the model is below this curve, it is not able to improve the *effective faithfulness* and it has a worse trade-off than the *control* operating at the same level of extractiveness.

For both Gigaword and Wikihow, Selector-ROC lies above the curve improving this trade-off. However, both the MLE baseline and Loss Truncation models get worse trade-off than the *control* operating at the same level of extractiveness. Similarly, we can obtain several models that lie above the curve for both Gigaword and Wikihow using Selector-F$_\beta$. The selector approach allows us to get better *effective faithfulness* at different points in the abstractiveness-extractiveness spectrum. The DAE based model is able to improve *effective faithfulness* on the Wikihow dataset, but not on the Gigaword dataset, indicating that the improvements are not consistent across datasets. Table 5 shows example summaries generated by the MLE baseline, Loss Truncation, DAE and the Selector-ROC models. We observe that selector model is able to generate summaries that are faithful to the original article while having more novel words and phrases in the generated summaries.

6 Related Work

There has been a lot of recent work in abstractive summarization showing that state-of-the-art systems suffer from generating inconsistent information with respect to the source article, despite their improved success in producing fluent summaries (Falke et al., 2019; Lux et al., 2020). Since word-overlap based metrics such as ROUGE have

low correlation with human scores of faithfulness (Kryscinski et al., 2019; Fabbri et al., 2020), there has been significant effort to develop automated metrics that can detect such errors (Zhou et al., 2021; Gabriel et al., 2021; Pagnoni et al., 2021a). For example, Falke et al. (2019), Maynez et al. (2020) and Goyal and Durrett (2020) have proposed to assess faithfulness using entailment models, where a faithful summary should be assigned a high entailment score with respect to the original article. Kryscinski et al. (2020) presented FactCC, a weakly-supervised BERT-based entailment model, by augmenting the dataset with artificial faithfulness errors. Durmus et al. (2020) and Wang et al. (2020) proposed question-answering based evaluation frameworks by automatically generating questions from the generated summary, and comparing the corresponding answers from both the source and the generated summary in order assess information consistency. Furthermore, several benchmarks have been proposed to evaluate the strengths and weaknesses of these evaluation metric (Gabriel et al., 2021; Pagnoni et al., 2021b).

Previous studies in faithfulness evaluation, however, has not accounted for the effect of extractiveness of the output summaries. As we show in this study, the extractiveness of the output is correlated with the faithfulness scores assigned by these automated metrics. Therefore, it is not clear whether the models with higher scores are better at abstraction, or extract more from the source article. We suggest that we need to account for this confounding factor in order to assess the real progress in building models that are better at abstraction. We note that there is concurrent work that also argues for accounting for extractiveness in assessing the faithfulness of models (Dreyer et al., 2021), however, unlike our work, they do they do not propose any mitigation for the faithfulness-abstractiveness trade-off.

Improving faithfulness of summarization systems is essential for deploying these systems in real-world scenarios, as such recent work has studied methods to improve the faithfulness of abstractive summarization systems (Zhao et al., 2020; Dong et al., 2020; Goyal and Durrett, 2021; Xu et al., 2020; Chen et al., 2021; Zhu et al., 2021). For example, Goyal and Durrett (2021) train summarization systems by modifying the training objective to maximize the likelihood of the subset of summary tokens that are considered faithful according to their factuality detection model. Zhao et al. (2020) specifically target hallucination of quantities in generated summaries, and train a verification model that they use to re-rank summaries such that summaries containing quantities consistent with the source article are up-ranked. Although these methods have shown improvements over the compared baselines, unlike our work, they do not measure the effective faithfulness taking extractiveness of the generated outputs into account.

7 Implications and Limitations

Recent studies that propose methods to improve faithfulness evaluate progress by conducting human evaluation on generated summaries and check whether the faithfulness scores are higher for their proposed system as compared to their baselines. We show that there is a strong relationship between the extractiveness and faithfulness of generated outputs (i.e., more extractive outputs tend to be more faithful), and therefore we cannot simply disregard extractiveness in faithfulness evaluation.

We propose that we should instead be measuring *effective faithfulness* and introduce a framework that takes into account the *faithfulness-abstractiveness trade-off curve* that is generated by training *control models* at different points in the abstractiveness spectrum. We demonstrate the importance of measuring *effective faithfulness* by showing that recently proposed methods that improve faithfulness over the MLE baseline fails to consistently improve over a simple *control* operating at the same level of abstractiveness.

We argue that measuring *effective faithfulness* is important since our goal is to build abstractive, faithful summarization systems. If the objective was to optimize for faithfulness alone, we could do so by simply building more extractive systems (such as the Q4 model we trained above).

Limitations. Note that this method relies on some diversity in the extractiveness of reference summaries, since we rely on sub-sampling to train models for the **control**. It is less likely to be effective for datasets with very little variation in the extractiveness of the generated summaries. However, in general, we see significantly more faithfulness problems for datasets with higher diversity of abstractiveness. Therefore, we suggest to account for the faithfulness-abstractiveness trade-off for such datasets in future work.

References

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

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

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.

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

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2021. Analyzing the abstractiveness-factuality tradeoff with nonlinear abstractiveness constraints.

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. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

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.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. Go-figure: A meta evaluation of factuality in summarization. In *ACL Findings*.

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.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Kang and Tatsunori 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, Online. Association for Computational Linguistics.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

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

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction

in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, 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.

Klaus-Michael Lux, Maya Sappelli, and Martha Larson. 2020. Truth or error? towards systematic analysis of factual errors in abstractive summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 1–10, 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.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021a. 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.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021b. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

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.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

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