FRSUM: Towards Faithful Abstractive Summarization via Enhancing Factual Robustness

Wenhao Wu[1]*, Wei Li[2], Jiachen Liu[2], Xinyan Xiao[2], Ziqiang Cao[3], Sujian Li[1]†, Hua Wu[2]
[1]Key Laboratory of Computational Linguistics, MOE, Peking University
[2]Baidu Inc., Beijing, China
[3]Institute of Artificial Intelligence, Soochow University, China
{waynewu,lisujian}@pku.edu.cn, {zqcao}@suda.edu.cn
{liwei85,xiaoxinyan,liujiachen,wu_hua}@baidu.com

Abstract

Despite being able to generate fluent and grammatical text, current Seq2Seq summarization models still suffering from the unfaithful generation problem. In this paper, we study the faithfulness of existing systems from a new perspective of factual robustness which is the ability to correctly generate factual information over adversarial unfaithful information. We first measure a model's factual robustness by its success rate to defend against adversarial attacks when generating factual information. The factual robustness analysis on a wide range of current systems shows its good consistency with human judgments on faithfulness. Inspired by these findings, we propose to improve the faithfulness of a model by enhancing its factual robustness. Specifically, we propose a novel training strategy, namely FRSUM, which teaches the model to defend against both explicit adversarial samples and implicit factual adversarial perturbations. Extensive automatic and human evaluation results show that FRSUM consistently improves the faithfulness of various Seq2Seq models, such as T5, BART.

1 Introduction

Abstractive summarization aims to produce fluent, informative, and faithful summaries for a given document. Benefiting from large-scale pre-training techniques, recent abstractive summarization systems are able to generate fluent and coherent summaries (Dong et al., 2019; Lewis et al., 2020; Xiao et al., 2020; Zhang et al., 2020a). However, challenges remain for this task. One of the most urgent requirements is to improve "faithfulness" or "factual consistency" of a model, which requires the generated text to be not only human-like but also faithful to the given document (Maynez et al., 2020). An earlier study observes nearly 30% of summaries suffer from this problem on the Gigawords dataset (Cao et al., 2018), while recent large-scale human evaluation concludes that 60% of summaries by several popular models contain at least one factual error on XSum (Pagnoni et al., 2021). These findings push the importance of improving faithfulness of summarization to the forefront of research.

Many recent studies focus on improving the faithfulness of summarization models, which can be mainly divided into three types. The first type modifies the model architecture to introduce pre-extracted guidance information as additional input (Cao et al., 2018; Dou et al., 2021; Zhu et al., 2021), while the second type relies on a post-editing module to correct the generated summaries (Dong et al., 2020; Chen et al., 2021). The last type takes advantages of auxiliary tasks like entailment (Li et al., 2018), and QA (Huang et al., 2020; Nan et al., 2021) on faithfulness. Different from previous studies, this work focuses on refining the training strategy of Seq2Seq models to improve faithfulness universally without involving any extra parameters, post-editing procedures and external auxiliary tasks.

In this paper, we study the faithfulness problem of Seq2Seq models from a new perspective of factual robustness, which is the robustness of generating factual information. We first define factual robustness as the model's ability to correctly generate factual information over adversarial unfaithful information. Following this definition, we analyze the factual robustness of a wide range of Seq2Seq models by measuring their success rate to defend against adversarial attacks when generating factual information. The analysis results (see Table 1) demonstrate good consistency between models' factual robustness and their faithfulness according to human judgments, and also reveal that current models are vulnerable to generating different types of unfaithful information. For example, the robust-

[*] Work is done during an internship at Baidu Inc.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3640 - 3654
December 7-11, 2022 ©2022 Association for Computational Linguistics

ness of generating numbers in the XSum dataset for most Seq2Seq models is very weak. Inspired by the findings above, we propose a novel faithful improvement training strategy, namely FRSUM, which improves a model's faithfulness by enhancing its factual robustness. Concretely, FRSUM teaches the model to defend against adversarial attacks by a novel factual adversarial loss, which constrains the model to generate correct information over the unfaithful adversarial samples. To further improve the generalization of FRSUM, we add factual adversarial perturbation to the training process which induces the model to generate unfaithful information. In this way, FRSUM not only requires the model to defend against explicit adversarial samples but also become insensitive to implicit adversarial perturbations. Thus, the model becomes more robust in generating factual spans, and generates fewer errors during inference. Moreover, FRSUM is adaptive to all Seq2Seq models.

Extensive experiments on several state-of-the-art Seq2Seq models demonstrate the effectiveness of FRSUM, which improves the faithfulness of various Seq2Seq models while maintaining their informativeness. Besides automatic evaluation, we also conduct fine-grained human evaluation to analyze different types of factual errors. The human evaluation results also show that FRSUM greatly reduces different types of factual errors. Especially, when applying to T5, our method reduces 17.5% and 49.1% of target factual errors on the XSum and CNN/DM datasets, respectively. Our contributions can be summarized as the following three points.

- We study the problem of unfaithful generation from a new perspective, factual robustness of Seq2Seq models, which is found consistent with faithfulness of summaries.
- We propose a new training method, FRSUM, which improves the factual robustness and faithfulness of a model by defending against both explicit and implicit adversarial attacks.
- Extensive automatic and human evaluations validate the effectiveness of FRSUM and also show that FRSUM greatly reduces different types of factual errors.

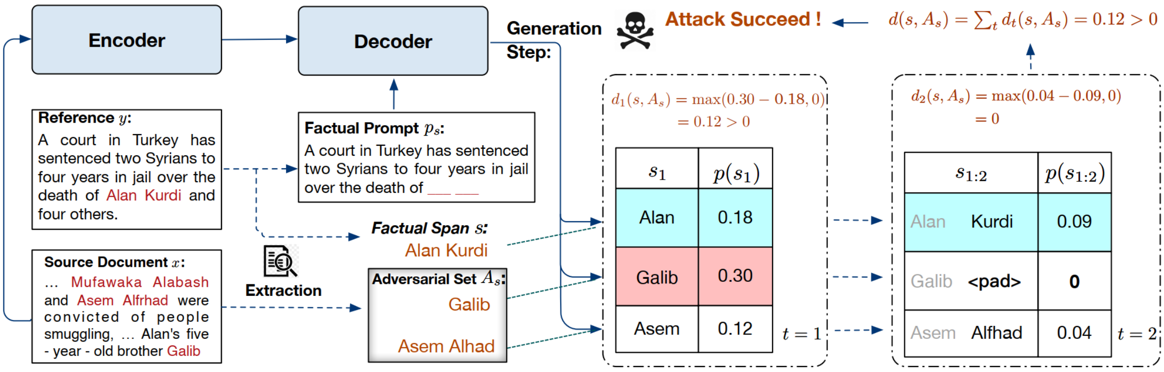
2 Related Work

2.1 Faithfulness of Summarization

Studies of faithfulness mainly focus on how to improve the faithfulness of an abstractive summarization model. Though it is challenging, some recent works propose various methods to study this problem, which can be summarized as follows. Some typical methods use pre-extracted information from input the document as additional input (Dou et al., 2021), like triplets (Cao et al., 2018), keywords (He et al., 2020), knowledge graphs (Huang et al., 2020; Zhu et al., 2021) or extractive summaries (Dou et al., 2021). These methods encourage the model to copy from the faithful guidance information. Another type of popular method focuses on designing a post-editing module, like a QA model (Dong et al., 2020), or a BART-based selection model (Chen et al., 2021). But these methods are much less time efficient during inference, thus hard to be used in real-world applications. Some other works apply Reinforcement Learning (RL) based methods, especially policy gradient, which utilize a variety of factual-relevant tasks for calculating rewards, such as information extraction (Zhang et al., 2020b), entailment (Li et al., 2018), QA (Huang et al., 2020; Nan et al., 2021). This type of methods suffer from the high-variance training of RL.

2.2 Adversarial Attacks for Text

Though deep neural networks (DNNs) have shown significant performance on various tasks, a series of studies have found that adversarial samples by adding imperceptible perturbations could easily fool DNNs (Szegedy et al., 2014; Goodfellow et al., 2015). These findings not only reveal potential security threats to DNN-based systems, but also show that training with adversarial attacks can enhance the robustness of a system (Carlini and Wagner, 2017). Recently, a large amount of studies focus on adversarial attacks for a variety of NLP tasks, such as text classification (Ebrahimi et al., 2018; Gil et al., 2019), question answering (Jia and Liang, 2017; Gan and Ng, 2019) and natural language inference (Minervini and Riedel, 2018; Li et al., 2020). Because of the discrete nature of language, these works mainly apply the methods of inserting, removing, or deleting different levels of text units (char, token, sentence) to build adversarial samples (Ren et al., 2019; Zang et al., 2020). Besides the aforementioned language understating tasks, some recent works also apply adversarial attacks on language generation. Cheng et al. (2019) applies adversarial attacks on both encoder and decoder to improve the performance of translation. Seq2Sick focuses on designing adversarial samples to attack



Reference y: A court in Turkey has sentenced two Syrians to four years in jail over the death of Alan Kurdi and four others.

Factual Prompt p_s: A court in Turkey has sentenced two Syrians to four years in jail over the death of ___ ___

Source Document x: ... Mufawaka Alabash and Asem Alfrhad were convicted of people smuggling, ... Alan's five - year - old brother Galib

s_1	$p(s_1)$
Alan	0.18
Galib	0.30
Asem	0.12

$s_{1:2}$	$p(s_{1:2})$
Alan Kurdi	0.09
Galib <pad>	0
Asem Alfhad	0.04

Figure 1: Procedure of an adversarial attack on a two-token entity span. After extracting a factual span s, a factual prompt p_s, and a set of corresponding adversarial samples A_s, we calculate the probability of generating s and spans in A_s given p_s. Based on the probability, we check whether this attack succeeds, according to Equation 4.

SeqSeq models for evaluating their robustness on informativeness (Cheng et al., 2020). Compared with previous works, we are the first to study the problem of unfaithful generation from the perspective of robustness.

3 Factual Robustness on Seq2Seqs

In this section, we introduce the definition and measurement of factual robustness. The factual robustness is defined as the ability of a Seq2Seq model to correctly generate factual information over adversarial unfaithful information. Formally, given a document x, a faithful summary y and a set of unfaithful adversarial samples A_s, a model with high factual robustness should satisfy:

$$p(y|x) > \max_{s^a \in A_s} p(s^a|x) \qquad (1)$$

where $p(y|x)$ is the probability of generating y given x, s^a is the adversarial sample in A_s. We adopt a process similar to adversarial attacks to measure the factual robustness of a Seq2Seq model. Extending from the conventional adversarial attack framework, we take the generation process of a factual information span as the target for attack. After constructing a set of adversarial samples, we check whether an attack succeeds by comparing the generation probabilities between the span and adversaries. We then define the measurement of factual robustness as the success rate of a model to defend against these attacks in a corpus. Following this definition, we measure the factual robustness of current models and analyze their relations with faithfulness.

3.1 Measurement of Factual Robustness

In this section, we measure the factual robustness of Seq2Seqs by adversarial attacks. Though adversarial attacks have been well-studied in classification tasks, it is still not straightforward to directly adapt them to text generation models. Different from attacking on a single label prediction in classification tasks, we consider the multi-step token predictions when generating a span of information.

Factual Span and Factual Prompt Given a document x and its reference with m tokes $y = \{y_1, y_2, \ldots, y_m\}$, we define a factual span s as the elementary unit of factual information, which can represent various types of facts. As the first study on factual robustness, we only analyze entity and number spans which are the most common types of information errors in existing summarization models. After extracting a span s, we define the prefix before s in y as *factual prompt* p_s, based on which the model should generate span s correctly.

Adversarial Sample s^a is a span that makes the information $[p_s, s^a]$ contradict the input x. It is used to attack the generation process of s. Previous study finds that intrinsic hallucinations are the most frequent factual errors in Seq2Seq models (Maynez et al., 2020). This kind of factual errors usually occur when the model confuses other information presented in the input document with the target information during generation. Thus, in this study, we construct *a set of adversarial samples* A_s by extracting entity and number spans from the source document x that are different from the target span s: $A_s = \{s^a | s^a \in x \& s^a \neq s\}$ to introduce intrinsic hallucinations.

System	XSum					CNN/DM				
	Mix%↓	Ent%↓	Num%↓	R-L↑	Incor%↓	Mix%↓	Ent%↓	Num%↓	R-L↑	Incor%↓
TransS2S	53.1	54.0	52.1	24.0	96.9	48.0	50.8	40.5	35.9	74.8
BERTSum	40.1	36.0	47.2	31.2	83.7	33.4	36.2	29.5	38.8	27.2
T5	37.3	33.2	43.4	33.1	82.0	37.5	40.2	31.9	40.2	26.7
BART	26.7	25.0	31.6	36.9	66.7	29.0	32.2	23.8	40.5	24.7
PEGASUS	22.4	20.0	29.0	39.1	60.7	28.3	29.6	22.2	40.5	13.3

Table 1: The factual robustness of different systems on CNN/DM and XSum datasets. Ent% and Num% are the success rate E of adversarial attack on entity and number spans, respectively. Mix% is the average success rate E of attacking both the number and entity spans. R-L is the abbreviation of ROUGE-L listed aside for reference. Incor% is incorrect ratio of generated summaries annotated by humans. The Pearson Correlation Coefficient and Spearman Correlation Coefficient between Mix% and Inroc% are 0.57 and 0.60, respectively.

Adversarial Attack We measure factual robustness by an adversarial attack process which utilizes the above adversarial samples. Specifically, given the input x and a factual prompt p_s, we apply adversarial attacks to auto-regressively generate s by using the adversarial samples in A_s. In every generation step, we check whether the model has the highest probability to generate the prefixes of s. Following conditional probability, in step t, the probability of generating the first t-token prefix of s ($t \leq |s|$, $|s|$ denotes the length of s) is:

$$p(s_{1:t}|p_s, x, \theta) = \prod_{i=1}^{t} p(s_i|s_{1:i-1}, p_s, x, \theta) \quad (2)$$

where s_i and $s_{1:i}$ are respectively the i-th token and first i-tokens of s, and θ denotes the model parameters. In the following, Eq. 2 is abbreviated as $p(s_{1:t})$. Based on it, we compare the probability of generating the target factual span $s_{1:t}$ and adversarial samples $s^a_{1:t}$ as:

$$d_t(s, A_s) = \max_{s^a \in A_s} (\max(p(s^a_{1:t}) - p(s_{1:t}), 0)) \quad (3)$$

which measures the tendency of generating unfaithful spans in adversarial samples over the factual span. To measure the full generation process, we average $d_t(s, A_s)$ of the total $|s|$ generation steps:

$$d(s, A_s) = \frac{1}{|s|} \sum_{t=1}^{|s|} d_t(s, A_s) \quad (4)$$

For the adversarial samples with different length $|s^a| \neq |s|$, s^a is truncated or padded to $|s|$, where the probability of generating the pad token is 0. In this way, each step we can compare the probability of generating prefixes of spans with the same length. If any adversarial sample has a higher probability, then $d(s, A_s) > 0$, indicating the success of this adversarial attack. An example of a successful adversarial attack is illustrated in Figure 1. In the first step, the model has a higher probability of generating the token " *Galib*" in adversarial samples instead of "*Alan*", so the adversarial attack succeeds.

Factual Robustness Following the definition above, we measure the factual robustness of a model via its success rate of attacks in a corpus. Given a test set D and a model with parameters θ, following Equation 4, the success rate of adversarial attack on D is calculated as:

$$E = \frac{\sum_{x,y \in D} \sum_{s \in y} \mathbb{1}[d(s, A_s) > 0]}{\sum_{y \in D} C_s(y)} \quad (5)$$

where $C_s(y)$ is the number of factual spans in the reference y, and $\mathbb{1}$ is the indicator function. Obviously, lower E indicates better factual robustness.

3.2 Factual Robustness and Faithfulness

Following Eq.5, we measure factual robustness of current SOTA Seq2Seq summarization systems and analyse their relations with faithfulness. We report both factual robustness and faithfulness of models on different datasets in Table 1. Details about these models and datasets are introduced in §5. We evaluate the factual robustness for two kinds of factual spans, i.e. entity and number. Their corresponding success rates of adversarial attacks are denoted as Ent% and Num%. Mix% is the average success rate of attacking all the entity and number spans in the reference summary y. Incor% denotes the ratio of unfaithful summaries judged by humans[1].

[1] Incor% annotation of T5 comes from § 5, while TransS2S and BERTSum come from Pagnoni et al. (2021), BART and PEGASUS come from Cao and Wang (2021).

From Mix% and Incor% reported in Table 1, we can conclude that factual robustness and faithfulness have good consistency: the more factually robust the model is (lower Mix%) the better faithfulness the generated summaries exhibit (lower Incor%). Specifically, the Pearson Correlation Coefficient and Spearman Correlation Coefficient between factual robustness (Mix%) and faithfulness (Incor%) are 0.57 and 0.60, respectively, which also show the great potential of utilizing factual robustness for faithfulness assessment. We also draw several other conclusions based on the results. Firstly, considering the simplicity of our adversarial samples, current systems are still vulnerable in factual robustness. Even the current SOTA models PEGASUS and BART fail to defend nearly 30% of the attacks. It can be further supposed that these models will have a lower factual robustness when defending against stronger adversarial samples. Secondly, a better pre-training strategy not only largely improves ROUGE scores but also improves the factual robustness and faithfulness, which is also confirmed by human evaluations (Maynez et al., 2020). Lastly, different types of factual spans perform differently regarding factual robustness. Generating numbers is more challenging in XSum than CNN/DM because it requires the model to comprehend and rewrite the numbers in the summaries rather than just copying spans contained in the input.

4 FRSUM

In the previous section, we introduce factual robustness and reveal its relation with faithfulness. We also discover that current systems are not robust enough in generating factual spans. Based on these findings, it is natural to improve a model's faithfulness by enhancing its factual robustness. Thus, we propose FRSUM, which is a training strategy to improve the faithfulness of Seq2Seq models by enhancing their factual robustness. FRSUM is composed of a factual adversarial loss and factual adversarial perturbation. Factual adversarial loss encourages the model to defend against explicit adversarial samples. Factual adversarial perturbations further apply implicit factual-relevant adversarial permutations to the previous procedure to enhance the factual robustness. We follow the notations in §3 to introduce FRSUM in detail.

FRSUM can be applied to all kinds of Seq2Seq models which are composed of an encoder and a decoder. Following the common Seq2Seq architecture, we apply Negative Log Likelihood (NLL) in the training process to generate fluent summaries. Given a document x and its reference y, the encoder first encodes input document $x = (x_1, x_2, \ldots, x_n)$ of length n into hidden representations h. After that, the decoder computes the NLL based on h and y:

$$\mathcal{L}_{nll}(\theta) = -\frac{1}{m}\sum_{t=1}^{m} \log p(y_t|y_{<t}, h, \theta) \quad (6)$$

4.1 Factual Adversarial Loss

In addition to NLL, we further propose the factual adversarial loss to enhance the model's factual robustness. As introduced in §3, we apply the success rate E of adversarial attack to measure a model's factual robustness. Similarly, we can also enhance factual robustness by optimizing E. Because Eq. 5 is discrete and intractable for gradient-based optimization, we apply the probability contrast between s and A_s (as in Eq. 3) for optimization instead. We first modify the probability contrast between two samples s and s^a by further adding a constant margin $\gamma > 0$ to adjust the degree of contrast:

$$d_t(s^a, s, \gamma) = \max(lp(s^a_{1:t}) - lp(s_{1:t}) + \gamma, 0)$$

where lp denotes the logarithm of the original p, t denotes the t-th generation step, consistent with previous sections. In this way, we encourage the model to generate faithful content over the adversaries by a margin in probability. Then, we expand the above pairwise probability contrast to a set of adversarial samples A_s and further compute the factual adversarial loss:

$$\mathcal{L}_{fa} = \frac{1}{C_s(y)}\sum_{s\in y}\frac{1}{|s|}\sum_{t=1}^{|s|}\max_{s^a\in A_s} d_t(s^a, s, \gamma) \quad (7)$$

4.2 Factual Adversarial Perturbation

Besides defending against explicit adversarial samples, we further apply implicit adversarial perturbations to enhance generalization of factual robustness (Madry et al., 2018). We propose factual adversarial perturbations and add them to the training process, which induce the model to have a higher probability to generate unfaithful information. In this way, FRSUM not only requires the model to defend against explicit adversarial samples but also become insensitive to implicit adversarial perturbations. Formally, the purpose of the perturbation is to disturb the generation of factual span s as much as possible. We measure the quality of generating s by its NLL given the factual prefix p_s:

$$l_s(\theta, h) = -\sum_{t=1}^{|s|} \log p(s_t | s_{1:t-1}, p_s, h, \theta) \quad (8)$$

For the simplicity of implementation, we add perturbation $\delta = [\delta_1 \ldots, \delta_n]$ on the encoded hidden states h. Following the definition of adversarial perturbation, the expected perturbation should satisfy the following condition:

$$\delta = \underset{\delta', ||\delta'|| \leq \epsilon}{\arg\max}\, l_s(\theta, h + \delta') \quad (9)$$

where the norm of δ' is constrained to be smaller than ϵ. We follow Goodfellow et al. (2015) to approximate δ by the first-order derivative of l_s, because the exact solution for δ is intractable in deep neural networks:

$$\delta = \nabla_h l_s(\theta, h) / \|\nabla_h l_s(\theta, h)\| \quad (10)$$

$$\widehat{h} = h + \tau * \delta \quad (11)$$

where $\widehat{h}$ is the hidden representation after perturbation, and τ is the update step. We then replace h with $\widehat{h}$ to predict the probability of generating s and s^a to compute $\mathcal{L}_{fa}^p$ following Eq.7, which is the perturbed version of $\mathcal{L}_{fa}$,

4.3 Training Procedure

The overall loss function of FRSUM is:

$$\mathcal{L} = \mathcal{L}_{nll} + \eta * \mathcal{L}_{fa}^p \quad (12)$$

where $\eta \in [0, 1]$ balances the NLL and factual adversarial loss. We gradually increase the difficulties of training by slowly increasing τ in Equation 11:

$$\tau = \min(\max((epoch - S), 0) * 0.1, 0.5) \quad (13)$$

where *epoch* is the number of current training epoch and S is the initial epoch to apply explicit adversarial perturbations. When $epoch > S$, τ gradually increase till the maximum of 0.5.

The whole training process is illustrated in Algorithm 1. For a given document-reference pair (x, y), we first extract and sample an entity or numeric span s from y and its corresponding adversarial set A_s from x (line 2-3), where $Sample(a, b)$ indicates sampling b samples from set a, $E()$ indicates the extraction of entity or number. After the model calculated $\mathcal{L}_{nll}$ (line 5-7), we add adversarial perturbations to h (line 9-10), where s_s and s_e are the start position and end position of s in y. After that, we apply $\hat{h}$ to calculate factual contrast loss $\mathcal{L}_{fc}^p$ based on the perturbated hidden state $\hat{h}$ (line 12-16). Finally, we use the final output loss $\mathcal{L}$ for training.

Algorithm 1: FRSUM

```
   Input  :Document $x$, Reference $y$, Entity and
           Number extractor E().
   Output :Training loss $\mathcal{L}$
1                                          ▷ Data Pre-processing
2  $s, p_s \leftarrow Sample(E(y), 1)$;
3  $A_s \leftarrow Sample(E(x) \setminus s, 10)$
4                                                   ▷ NLL Loss
5  $h = Encoder(x)$
6  $P_{tgt} = Decoder(h, y) = [p_1, p_2, \ldots, p_m]$;
7  $\mathcal{L}_{nll} = -\frac{1}{m} \sum_{i=1}^{m} \log P_{tgt}[i]$;
8                               ▷ Factual Relevant Permutation
9  $l_s(\theta, h) = -\frac{1}{|s|} \sum_{i=s_s}^{s_e} \log P_{tgt}[i]$
10 $\hat{h} = h + \epsilon * \nabla_h l_s / |\nabla_h l_s|$
11                                    ▷ Factual Contrast Loss
12 $p(f) = Decoder(\hat{h}, [p, f])$
13 for $s^a$ in $A_s$ do
14     $p(s^a) = Decoder(\hat{h}, [p_s, s^a])$
15 end
16 $\mathcal{L}_{fc}^p \leftarrow$ Eq.7 with $p(s)$,$\{p(s^a) | a^a \in A_s\}$
17                                              ▷ Output Loss
18 $\mathcal{L} = \mathcal{L}_{nll} + \eta * \mathcal{L}_{fc}^p$
```

5 Experiment Setup

5.1 Datasets

CNN/DM CNN/DM is a news dataset with multi-sentence summaries. CNN/DM contains news articles and associated highlights, which are used as a multi-sentence summary. We used the standard splits of Hermann et al. for training, validation, and testing (90,266/1,220/1,093 CNN documents and 196,961/12,148/10,397 DailyMail documents). We used pre-processed version from See et al., and the input documents were truncated to 512 tokens.

XSum XSum (Narayan et al., 2018) is a news dataset for extreme summarization, which requires the model to summarize a news document with only one sentence summary. We used the splits of Narayan et al. (2018) for training, validation, and testing (204,045/11,332/11,334) and followed the pre-processing introduced in their work. Input documents were truncated to 512 tokens.

	CC	SC	R-1	R-2	R-L
TransS2S	83.27	79.10	39.30	17.27	35.89
Split Encoders	73.11	-	38.83	16.51	35.71
Fact Correction	82.82	-	39.87	17.50	36.80
FRSUM	**83.75**†	**79.97**†	**41.25**	**18.96**	**37.99**
BERTSUM	75.73	71.66	41.72	19.39	38.76
Split Encoders	76.43	-	39.78	17.87	37.01
Fact Correction	78.69	-	41.13	18.58	38.04
FRSUM	**77.18**†	**72.21**	41.59	19.03	38.66
BART	80.66	78.66	**43.88**	**20.93**	**40.57**
ContrastSel	83.23	74.03	42.66	19.82	39.34
CLIFF	78.13	77.59	43.92	20.95	40.60
FRSUM	**83.24**†	**80.40**†	43.54	20.61	40.19
T5	75.32	74.76	**43.24**	**20.65**	**40.18**
ContrastSel	73.55	72.73	43.05	20.45	40.00
CLIFF	74.82	73.31	42.72	19.96	39.41
FRSUM	**76.43**†	**74.82**	42.92	20.24	39.69

Table 2: Evaluation results of FRSUM on CNN/DM. †: FactCC (**CC**) or SC **SC** significantly better than models trained only with NLL (in gray) ($p < 0.05$) in T-test.

	CC	SC	R-1	R-2	R-L
TranS2S	24.15	22.50	30.51	10.30	24.20
Split Encoders	24.78	-	29.45	9.59	23.40
Fact Correction	25.75	-	36.24	14.37	29.22
FRSUM	**28.47**†	**24.04**†	**31.38**	**10.89**	**25.01**
BERTSUM	23.81	23.72	38.76	16.33	31.15
Split Encoders	24.19	-	34.22	13.76	27.86
Fact Correction	**25.08**	-	36.24	14.37	29.22
FRSUM	24.03	**23.80**	**38.79**	**16.46**	**31.22**
BART	23.64	25.72	**45.20**	**21.90**	**36.88**
ContrastSel	24.89	23.59	44.54	21.23	36.28
CLIFF	23.51	**25.91**	44.63	21.39	36.43
FRSUM	**25.52**†	25.80	44.75	21.66	36.76
T5	23.93	22.97	41.15	18.18	33.10
ContrastSel	26.22	21.58	40.87	17.71	32.68
CLIFF	**26.72**	22.27	39.48	16.26	31.40
FRSUM	24.86†	**23.12**	**41.42**	**18.50**	**33.48**

Table 3: Evaluation results of FRSUM on XSum.

5.2 Automatic Metric

Informative Metric We evaluate the informativeness of generated summaries using ROUGE F_1 (Lin, 2004). Specifically, we use ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L).

Factual Metric We evaluate the faithfulness of the generated summaries by **FactCC (CC)** (Kryscinski et al., 2020). Recent large-scale human evaluation validates that FactCC correlates well with human judgments on both CNN/DM and XSum (Pagnoni et al., 2021). We also apply another recent factual metric **SummaC (SC)**, which achieves the state-of-the-art performance on summary inconsistency detection benchmark (Laban et al., 2021).

5.3 Baselines

We select various Seq2Seq models as backbones, based on which we separately compare FRSUM with different faithful improvement methods.

Seq2Seq Models We evaluate FRSUM on extensive baseline systems, especially SOTA pre-trained models. For non-pretrained models, we select the vanilla Transformer-based (Vaswani et al., 2017) Seq2Seq (TranS2S) as the representative. For pre-trained models, we select the following: partially pre-trained model, BertSum (Liu and Lapata, 2019); unified pre-trained model for both language understanding and generation, T5 (Raffel et al., 2019); pre-trained model for language generation tasks, BART (Lewis et al., 2020).

Faithful Improvement We also compare against other recent faithfulness improvement methods: **Split Encoders** (Shah et al., 2020), a two-encoder pointer generator; **Fact Correction** (Dong et al., 2020), a QA-based based model that correct the errors in the summary; **ContrastSel** (Chen et al., 2021), a BART-based classifier that selects faithful summaries in beam search; **CLIFF** (Cao and Wang, 2021), a contrastive learning based training method utilizing various synthetic augmentation samples. Please refer to Appendix A for implementation details.

6 Results

Because FRSUM focuses on faithfulness, we expect improvements on factual metrics without harming the performance of informative metrics. We select the T5 model for an ablation study and human evaluations because it is a widely used model with a relatively moderate factual robustness $E\%$.

6.1 Automatic Evaluation

The experimental results on CNN/DM and XSum datasets are reported in Table 2 and 3. **FRSUM** in the last column of each block reports the performance of the baseline further trained with FRSUM. FRSUM significantly improves the faithfulness of all Seq2Seq baselines and also outperforms all other recent faithfulness improving methods

Dataset	XSum				CNN/DM			
	CC↑	**SC**↑	$E\%$ ↓	R-L	**CC**↑	**SC**↑	$E\%$ ↓	R-L
TranS2S	24.2	22.5	53.1	24.2	83.3	79.1	48.0	35.9
+FRSUM	**28.5**	**24.0**	**49.6**	**25.0**	**83.8**	**80.0**	**43.3**	**38.0**
BertSum	23.8	24.0	40.1	31.2	75.7	71.7	33.4	38.8
+FRSUM	**24.0**	23.8	**38.5**	**31.2**	**77.1**	**72.2**	**31.0**	38.7
BART	23.6	25.7	26.7	36.9	80.7	78.7	29.0	40.6
+FRSUM	**25.5**	**25.8**	**24.3**	36.8	**83.2**	**80.4**	**27.5**	40.2
T5	23.9	23.0	37.3	33.1	75.3	74.8	37.5	40.1
+FRSUM	**24.9**	**23.1**	**35.7**	**33.5**	**76.4**	74.8	**36.4**	39.7
w/o per	24.2	22.8	36.2	33.2	74.0	**75.0**	37.2	40.0
w/o fa	24.4	22.6	36.3	33.3	74.3	72.9	37.0	39.8

Table 4: **Factual Robustness Analysis** and **Ablation Study** of FRSUM. $E\%$ denotes the measurements of factual robustness. **per** and **fa** refer to factual adversarial permutation and factual adversarial loss.

most of the time, achieving the best CC scores on 6 of 8 settings, the best SC scores on 7 of 8 settings. Although in some specific cases, some baseline methods such as T5-based CLIFF and BERTSUM-based Fact Correction on XSUM have higher FactCC scores than FRSUM, their ROUGE scores drop significantly. By contrast, FRSUM maintains the informativeness of baselines well and even improves the ROUGE scores of several baseline methods, for example, FRSUM improves ROUGE scores of 3 baselines on the XSum dataset. Overall, compared with other faithfulness improving methods, *FRSUM extensively demonstrates its superiority on improving faithfulness while preserving informativeness*.

Factual Robustness Analysis We further analyze the faithfulness of FRSUM considering its effects on factual robustness in Table 4, where $E\%$ in the table reports the factual robustness of systems. Concretely, $E\%$ equals to $Mix\%$ in Table 1. According to the results, we can conclude that FRSUM consistently improves the CC score and almost all the SC of all baseline methods while reducing $E\%$, and thus improves faithfulness.

Ablation Study The results of ablation study are reported in the last two rows in Table 4. **w/o permut** represents removing the factual adversarial perturbation of FRSUM, and **w/o fa** represents removing the factual adversarial loss and applying factual adversarial permutations on NLL. After removing factual adversarial permutations or factual adversarial loss, FRSUM decreases in CC, SC and increases on $E\%$. Thus, we conclude that these two mechanisms can work separately and combining them further improves the faithfulness.

6.2 Human Evaluation

We further conduct human evaluations to assess the effectiveness of FRSUM. Instead of comparing systems pairwise for faithfulness like previous studies, we report the exact number of different types of factual errors. We adopt the linguistically grounded typology of factual errors from Frank (Pagnoni et al., 2021). According to Frank, we divide factual errors into 5 types: Entity Error (EntE), Circumstance Error (CircE), Out of Article Error (OutE), Predicate Error (PredE), and Other Error (OtherE). EntE and OutE relate to entity errors, and CircE mainly relates to numeric errors. EntE captures entity errors that are contained in the input, while OutE captures entity errors that are not contained in the input.For informativeness evaluations, we apply a pairwise comparison between FRSUM and the original T5. We invite two professional annotators and randomly select 100 samples from both XSum and CNN/DM test sets for evaluation.

We report the average results in Table 5, where "Inf." denotes the ratio of summaries that have a better informativeness than the other systems. From the number of total errors we can see that FRSUM reduces factual errors of T5 in both datasets by 10.6% and 41.7%, respectively. Regarding specific error types, FRSUM substantially reduces EntE and CircE, which are the target types of factual spans for adversarial attacks. In total, FRSUM reduces the number of target errors by 18.4% and 49.1% on XSum and CNN/DM, respectively. We also notice that models generate a large number of OutEs on XSum. This is because the XSum dataset itself contains a large number of OutEs in the reference summary while FRSUM is not designed to overcome such noise (Gehrmann et al., 2021).

7 Conclusions and Future Work

In this paper, we study the faithfulness of abstractive summarization from the new perspective of factual robustness. We propose a novel adversarial attack method to measure and analyze the factual robustness of current Seq2Seq models. Furthermore, we propose FRSUM, a faithful improvement training strategy by enhancing the factual robust-

Model	**EntE**	**CircE**	OutE	PredE	OtherE	#Target	#Total	Inf%↑
T5	24.5	27.0	**35.0**	16.0	1.0	51.5	103.5	30.0
CLIFF	21.0	24.0	38.0	13.0	1.0	45.0	97.0	26.5
ContrastSel	20.0	25.5	40.0	11.0	0.0	45.5	96.5	24.5
FRSUM	**20.3**	**22.5**	35.5	14.5	1.5	**41.0**(18.4% ↓)	**92.5**(10.6% ↓)	**34.0**

(a) XSum

Model	**EntE**	**CircE**	OutE	PredE	OtherE	#Target	#Total	Inf%↑
T5	13.0	14.5	0.5	1.5	0.5	27.5	30.0	40.0
CLIFF	10.0	8.0	1.0	0.5	0.5	18.0	20.0	26.0
ContrastSel	12.0	10	0.5	0.0	0.0	22.0	22.5	22.5
FRSUM	**6.0**	**8.0**	2.0	1.5	0.0	**14.0** (49.1% ↓)	**17.5**(41.7% ↓)	**42.5**

(b) CNN/DM

Table 5: Human evaluation results on XSum and CNN/DM, where the kappa scores are 0.45 and 0.70, respectively. The second to sixth columns report the number of each type of factual errors. **#Total** and **#Target** report the number of total types and target types of factual errors, respectively. **Inf%** denotes the frequency of the system summary ranked first in informativeness. All the numbers are the average scores of two annotators. Brackets in #Target and #Total report the percentage of relative error decrease of FRSUM over BART.

ness of a model to improves its faithfulness.

Limitations

As the first study on factual robustness, we only analyze entity and number spans which are the most common types of information errors in existing summarization models. Future works can study more complicated factual errors such as relation errors.

Acknowledgments

We thank the anonymous reviewers for their helpful comments on this paper. This work was partially supported by National Key Research and Development Project (2020AAA0109703), National Natural Science Foundation of China (61876009), and National Social Science Foundation Project of China (21&ZD287).

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference*, pages 86–90. Morgan Kaufmann Publishers / ACL.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6633–6649. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

Nicholas Carlini and David A. Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5935–5941. Association for Computational Linguistics.

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3601–3608. AAAI Press.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4324–4333. 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. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

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

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

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36. Association for Computational Linguistics.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6065–6075. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, 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, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. 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. *CoRR*, abs/2102.01672.

Yotam Gil, Yoav Chai, Or Gorodissky, and Jonathan Berant. 2019. White-to-black: Efficient distillation of black-box adversarial attacks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1373–1379. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Fatema Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *CoRR*, abs/2012.04281.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5094–5107. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2021–2031. 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 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics.

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

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.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1430–1441. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6193–6202. Association for Computational Linguistics.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3728–3738. Association for Computational Linguistics.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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.

Susan Weber McRoy. 2000. Gillian brown, speakers, listeners, and communication: Explorations in discourse analysis. *User Model. User Adapt. Interact.*, 10(4):309–313.

Pasquale Minervini and Sebastian Riedel. 2018. Adversarially regularising neural NLI models to integrate logical background knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 65–74. Association for Computational Linguistics.

Feng Nan, Cícero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen R. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6881–6894. Association for Computational Linguistics.

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

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

Martha Palmer, Paul R. Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguistics*, 31(1):71–106.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Darsh J. Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. In *The*

Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8791–8798. AAAI Press.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Dongling Xiao, Han Zhang, Yu-Kun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3997–4003. ijcai.org.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6066–6080. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis Langlotz. 2020b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5108–5120. 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.

Model	Dataset	Training Steps	Learning Rate	Batch Size
T5	XSum	50k	1e-2	128
	CNN/DM	50k	1e-2	128
BART	XSum	20k	5e-5	64
	CNN/DM	15k	5e-5	128
PEGASUS	XSum	80k	1e-4	256
	CNN/DM	170k	5e-5	256

Table 6: Parameter settings of pre-train based models used in our experiments

A Hyper-parameter Details

For TransS2S, we set the number of both transformer encoder and decoder layers to 6 and the hidden state dimension to 512. For other pre-training based models, we follow their original parameters for training. We apply the base-version of T5 and large-version of BART and PEGASUS. The detailed training settings of all the baseline models are set in Table 6. We apply beam search for inference. During inference, for the XSum dataset, we set beam size to 6, alpha to 0.90, maximum length to 100, maximum length to 10; for CNN/DM dataset, we set beam size to 5, alpha to 0.95, maximum length to 150, maximum length to 30. For a fair competition, we report the results of CLIFF trained with negative samples constructed by entity swap.

For FRSUM, we apply Spacy for extracting entities and numbers. In the training process of factual adversarial loss, we randomly sample one s in y for optimization, which we find easier for training. And we also find a larger size of A_s leads to better performance. Thus in practice, we constrain the maximum size of A_s to 10 due to memory constraints. For time efficiency, we trained the model with FRSUM on the checkpoint when the model is close to coverage. η is set to 0.3, λ is set to 0.05 and S is the second epoch that the model starts to apply FRSUM for training. All the models are trained on 8 Nvidia V100 GPUS.

B Human Evaluation Details

Typology of Factual Errors Recently, Pagnoni et al. (2021) proposes a typology of factual errors which is theoretically grounded in frame semantics (Baker et al., 1998; Palmer et al., 2005), and linguistic discourse analysis (McRoy, 2000). This typology divided factual errors into 7 different categories including Circumstance Error (CircE), Entity Error (EntE), Out of Article Error (OutE), PredE (Relation Error), Coreference Error (CorefE), Discourse Link Error (LinkE), Grammatical Error (Gram-

	Category	Description	Example
CircE	Circumstance Error	The additional information (like location or time) specifying the circumstance around a predicate is wrong.	A 22-year-old teenager has been charged in connection with a serious assault in Bridge Street.
EntE	Entity Error	The primary arguments (or their attributes) of the predicate are wrong.	A teenager has been charged in connection with a serious assault in Aberdeen Sheriff Court.
OutE	Out of Article Error	The statement contains information not present in the source article.	A teenager has been charged in connection with a serious assault in London.
PredE	Relation Error	The predicate in the summary statement is inconsistent with the source article.	A teenager is not charged in connection with a serious assault in Bridge Street.
OtherE	Other Error	Other factual errors like Grammatical Error, Discourse Error.	A teenager has been charged in connect with a serious assault in Bridge Street. (GrammarE)

Table 7: Typology of factual errors in out human evaluation. Original text from the XSum dataset for the examples:*The 22-year - old man needed hospital treatment after the incident on Bridge Street on New Year's Day. Police Scotland said a 15-year - old boy had been charged. The teenager is expected to appear at Aberdeen Sheriff Court.*

merE). Because CorefE, LinkE, and GrammerE seldomly appear in generated summaries, in our study, we categorize them jointly as OtherE. The definitions and examples of typology of factual errors are illustrated in Table 7.

Annotation Details Each annotator is first trained to recognize and classify factual errors into a certain category by comparing summaries with the input documents. A summary may contain more than one factual error. During annotation, each annotator is given a document with two generated summaries from T5 and FRSUM, respectively. After annotating all the factual errors in these summaries, the annotator also needs to judge which summary is more or equally informative.

C Case Study

We show some cases to demonstrate our human evaluation and the effectiveness of FRSUM in Table 8 and Table 9 on XSum and CNN/DM datasets, respectively. From Document 1 and Document 2, we illustrate how FRSUM reduces CircE and EntE on XSum. Document 3 illustrates a special case where the Baseline model generates two errors, OutE and EntE. Notice that its gold reference also contains OutE, we can infer that the generated OutE is mainly caused by the unfaithful reference in training. Applying FRSUM on baseline reduces the EntE error but can not reduce the OutE. Table 9 illustrates FRSUM reduces numeric errors (CircE) including date, frequency and score, of 3 examples from CNN/DM.

XSum Human Evaluation Cases	
Document 1	The animal had been shot twice in the shoulder and once in its left back leg, which vets had to amputate.The charity said the one-year-old cat was "incredibly lucky" to survive.Last year the Scottish government held a consultation on licensing air weapons, but a majority of responders opposed the plan.One-year-old Teenie was found injured by her owner Sarah Nisbett in NiddryView, Winchburgh, at about 16:30 on Friday 14 March and taken to the Scottish SPCA.Mrs Nisbett said the cat was now having to learn how to walk again."The gun that was used must have some power because the pellet actually went through her back leg, that's why it was so badly damaged," she said."She's now learning how to hop around the house, it's terrible."The fact that it was three shots is crazy. We live in a housing estate and there are lots of kids. That just makes it worse because any of them could have been hit in the crossfire."She added: "There's some sick people out there, hopefully somebody will know who's done this and let the police or the Scottish SPCA know."Scottish SPCA Ch Supt Mike Flynn said: "Teenie's owners are understandably very upset and keen for us to find the callous person responsible to ensure no more cats come to harm."This is an alarming incident which only highlights why the Scottish government should implement the licensing of airguns as a matter of urgency."He added: "The new licensing regime should ensure that only those with a lawful reason are allowed to possess such a dangerous weapon. It will also help the police trace anyone using an air gun irresponsibly."...
Baseline	The Scottish SPCA has appealed for information after a cat was shot twice in the leg in West Lothian. (CircE)
+FRSUM	The Scottish SPCA has appealed for information after a cat was shot three times in a crossfire.
Document 2	It comes in a shake-up of UK military buildings and resettling of regiments.Brecon and Radnorshire Conservative MP Chris Davies condemned the closure, saying there had been a barracks in Brecon since 1805, home to troops who fought the Zulus at Rorke's Drift."This decision is abhorrent and I shall be fighting it every step of the way," he said."The government has a great deal of questions to answer over why it is proposing to close a well-loved and historic barracks in a vitally important military town."Brecon Barracks has served our country with distinction over its long history, with soldiers from the site fighting in every conflict since the early 19th century."This decision shows a blatant lack of respect for that history."Mr Davies said he was launching a petition against the decision, saying the Brecon area had some of the highest unemployment levels in Wales.He also hoped the closure would not damage the town's "thriving" military tourism industry.Brecon barracks has about 85 civilian staff and 90 military but it is not thought jobs are at risk.Mr Davies said he understood the nearby Sennybridge training ground and infantry school at Dering Lines would not be affected.Defence Secretary Sir Michael Fallon told the Commons on Monday the reorganisation in Wales would see a specialist light infantry centre created at St Athan, Vale of Glamorgan. Cawdor Barracks, Pembrokeshire - whose closure was previously announced in 2013 - will now shut in 2024, while a storage depot at Sennybridge will go in 2025.Responding for Labour, Shadow Defence Secretary Nia Griffith, MP for Llanelli, said the ministry was "right to restructure its estate".But she warned closing bases would affect the livelihoods of many people who would face "gnawing uncertainty" over their future.
Baseline	The government's decision to close military bases in Powys is " abhorrent ", an MP has said.(EntE)
+FRSUM	Plans to close the Brecon Barracks in Powys have been described as " abhorrent ".
Document 3	Jung won aboard Sam, who was a late replacement when Fischertakinou contracted an infection in July.France's Astier Nicolas took silver and American Phillip Dutton won bronze as GB's William Fox-Pitt finished 12th.Fox-Pitt, 47, was competing just 10 months after being placed in an induced coma following a fall.The three-time Olympic medallist, aboard Chilli Morning, produced a faultless performance in Tuesday's final show-jumping phase.But the former world number one's medal bid had already been ruined by a disappointing performance in the cross-country phase on Monday.He led after the dressage phase, but dropped to 21st after incurring several time penalties in the cross country.Ireland's Jonty Evans finished ninth on Cooley Rorkes Drift.Why not come along, meet and ride Henry the mechanical horse at some of the Official Team GB fan parks during the Rio Olympics?Find out how to get into equestrian with our special guide.Subscribe to the BBC Sport newsletter to get our pick of news, features and video sent to your inbox.
Gold	Germany's Michael Jung closed in on a £240,000 bonus prize as he secured a dominant lead to take into the final day of Badminton Horse Trials. (OutE)
Baseline	Germany's Sam Jung won Olympic gold in the equestrian with victory in the dressage phase on the back of a rider ruled out by illness. (OutE, EntE)
+FRSUM	South Africa's won Olympic gold in the equestrian event at Rio 2016 as Greece's Georgios Fischertakinou was hampered by an infection. (OutE)

Table 8: Three samples from human evaluations on XSum dataset.

CNN/DM Human Evaluation Cases	
Document 4	Lewis Hamilton has conceded to feeling more powerful now than at any stage in his F1 career. It is an ominous warning from a man who has won nine of the last 11 grands prix, been on pole at the last four, and who already holds a 27-point cushion in the drivers’ standings. It is no wonder after winning in Bahrain, when Hamilton stepped out of his Mercedes, he immediately stood on top of it and pretended to smack an imaginary baseball out of the circuit. Lewis Hamilton stands on his Mercedes after winning the Bahrain Grand Prix It was another ’home run’ performance from Hamilton, a man who claims he is a perfectionist, and who appears to be driving as close to perfection as can possibly be achieved in the sport. It led to the suggestion that perhaps he was feeling unbeatable, to which he replied: ’I don’t know what the feeling of being unbeatable is. ’I know I feel very powerful in this car with the package we have, and I feel I’m able to get everything from it. ’I also feel more comfortable in this car than I did in the one last year, and I feel within myself, with whatever approach I have, the power is greater than ever in terms of strength in the car. ’It’s like ’yes’. I put the car there and it goes there, I get the exit. ’Naturally you always feel there are things you can still improve on, but it’s hard to always pinpoint what those are and to tell the team what to focus on. Hamilton is congratulated by Ferrari rival Kimi Raikkonen on Sunday ’That’s what a team leader has to do, and, along with Nico (Rosberg), I’m trying to guide the team, to tell them where to go next.’ ...
Baseline	"Lewis Hamilton won the Bahrain Grand Prix on Sunday. The 30-year - old has won nine of the last 11 grands prix. He has been on pole at the last four and holds a 27-point cushion. Hamilton claims he is a perfectionist. " (CircE)
+FRSUM	"Lewis Hamilton has conceded to feeling more powerful now than at any stage in his F1 career. The 30-year - old has won nine of the last 11 grands prix and been on pole at the last four. Hamilton holds a 27-point cushion in the drivers’standings. "
Document 5	Jermain Defoe says his stunning Wear-Tyne derby winner is justification for his decision to quit MLS and return to the Premier League. The former FC Toronto striker had scored just twice for struggling Sunderland since arriving on £80,000-per-week wages in January. But Defoe was the hero on Sunday as his 20-yard volley proved enough to secure a fifth straight victory for the Black Cats over Newcastle. Sunderland striker Jermain Defoe believes his stunning volley against Newcastle has proven his worth Defoe’s superb first-half strike was enough to secure a 1-0 win for Sunderland in the Wear-Tyne derby Newcastle goalkeeper Tim Krul was completely helpless as Defoe’s shot found its way into the top corner The 32-year-old was overcome with emotion in the wake of his brilliant blast, and admits the joy it brought to a sold-out Stadium of Light was too much to take in. ...
Baseline	Sunderland beat Newcastle 1 - 0 in the Wear - Tyne derby on Sunday. Jermain Defoe scored a stunning first - half volley for the Black Cats. The former FC Toronto striker had scored just twice for the club. "(CircE)
+FRSUM	"Sunderland beat Newcastle 1 - 0 in the Wear - Tyne derby on Sunday. Jermain Defoe scored a stunning volley in the first half. Defoe had scored just twice for struggling Sunderland since January. "
Document 6	Former Valencia striker Aritz Aduriz denied his old team victory with a last-gasp equaliser for Athletic Bilbao at San Mames Stadium. Aduriz pounced in the 90th minute to secure a 1-1 draw after Valencia had been reduced to 10 men. Nicolas Otamendi had harshly received a straight red card eight minutes earlier for a high challenge, and Valencia were unable to hold out in his absence. Athletic Bilbao Aritz Aduriz scored a 90th minute equaliser to deny his former club Valencia victory Substitute Rodrigo De Paul celebrates after coming on to give Valencia the lead Substitute Rodrigo De Paul had given the visitors a deserved lead 16 minutes into the second period, but even though they were unable to claim maximum points, a draw took them above Sevilla into fourth place in their pursuit of possible Champions League qualification. Getafe bolstered their hopes of securing Primera Division status for next season after moving 10 points clear of the bottom three by edging to a 1-0 win at struggling Elche. Fredy Hinestroza scored the only goal of the game with only five minutes remaining in a rare attack for Getafe, after the hosts had spurned a series of chances in the first half. Getafe celebrate after taking a step closer to survival by defeating Elche through Fredy Hinestroza’s strike Ecuadorian Felipe Caicedo scored twice for Espanyol in the 3-0 defeat of Villarreal Elsewhere, Espanyol extended Villarreal’s winless streak to six games in all competitions as two goals from Felipe Caicedo handed the Catalans a 3-0 win at El Madrigal. Caicedo had a slice of fortune in opening the scoring as Javi Lopez’s miscued header deflected off the Ecuadorian and he then thumped home his second of the evening after the break at Sergio Asenjo’s near post. Asenjo had a night to forget as he also spilled Sergio Garcia’s cross 11 minutes from time to allow Victor Sanchez to roll the ball into an empty net to complete Villarreal’s misery.
Baseline	"Aritz Aduriz scored a 90th minute equaliser for Athletic Bilbao. The former Valencia striker denied his former club’s 1-1 victory. Getafe moved up to fourth in the table after a 1 - 0 win at Elche. Felipe Caicedo scored twice for Espanyol in the 3 - 0 defeat of Villarreal. " (CircE)
+FRSUM	"Aritz Aduriz scored a 90th minute equaliser for Athletic Bilbao. Valencia were reduced to 10 men after Nicolas Otamendi was sent off. Getafe moved 10 points clear of the bottom three with a 1 - 0 win at Elche. Felipe Caicedo scored twice for Espanyol against Villarreal. "

Table 9: Three samples from human evaluations on CNN/DM dataset.