Precisely the Point: Adversarial Augmentations for Faithful and Informative Text Generation

Wenhao Wu[1][*], **Wei Li**[2], **Jiachen Liu**[2], **Xinyan Xiao**[2], **Sujian Li**[1][†], **Yajuan Lyu**[2]
[1]Key Laboratory of Computational Linguistics, MOE, Peking University
[2]Baidu Inc., Beijing, China
{waynewu,lisujian}@pku.edu.cn
{liwei85,liujiachen,xiaoxinyan, lvyajuan}@baidu.com

Abstract

Though model robustness has been extensively studied in language understanding, the robustness of Seq2Seq generation remains understudied. In this paper, we conduct the first quantitative analysis on the robustness of pre-trained Seq2Seq models. We find that even current SOTA pre-trained Seq2Seq model (BART) is still vulnerable, which leads to significant degeneration in faithfulness and informativeness for text generation tasks. This motivated us to further propose a novel adversarial augmentation framework, namely AdvSeq, for generally improving faithfulness and informativeness of Seq2Seq models via enhancing their robustness. AdvSeq automatically constructs two types of adversarial augmentations during training, including implicit adversarial samples by perturbing word representations and explicit adversarial samples by word swapping, both of which effectively improve Seq2Seq robustness. Extensive experiments on three popular text generation tasks demonstrate that AdvSeq significantly improves both the faithfulness and informativeness of Seq2Seq generation under both automatic and human evaluation settings.

Original Input x: Police said the incident happened near the village of Tockwith. North Yorkshire Police said it was believed the aeroplane had suffered engine failure. Sgt Andy Graham, who attended the crash, praised the "great piloting skills resulting in no injuries".

Output from x: A light aircraft has crashed into a field in Yorkshire, but no-one was injured.

Adversarial Input x': Police said the incident happened near the village of Tockwith. North Yorkshire Police said it was believed the aeroplanes had suffered engine failure. Sgt Andy Graham, who attended the crash, praised the "great piloting skills resulting in no accident".

Output from x': Two aeroplanes have crashed in North Yorkshire, with one of them landing safely on its roof.

Table 1: A sample of the XSum dataset for text summarization generated by fine-tuned BART, where minor perturbations (words in blue) cause low-informative and unfaithful generation (span in red).

1 Introduction

Recently, text generation has made significant progress thanks to the development of pre-trained sequence-to-sequence (Seq2Seq) models (Lewis et al., 2020; Zhang et al., 2020; Roller et al., 2021). Despite being able to generate fluent and grammatical text, current state-of-the-art models tend to generate low-informative and unfaithful outputs that hallucinated with the given inputs (Welleck et al., 2019; Maynez et al., 2020). One major reason is that Seq2Seq models, generally trained with Negative Log-Likelihood (NLL), are still not robust enough to handle perturbations in the input (Kang and Hashimoto, 2020). This vulnerability can lead to significant degeneration of the generalization performance for text generation tasks (Cheng et al., 2019). Table 1 provides a typical example to show the vulnerability of the model to perturbations. We can see that the words "aeroplane" and "injuries" in the original text are revised as "aeroplanes" and "accident" respectively, which do not alter the meaning of the input. However, the output generated by BART obviously hallucinates the input.

The model robustness problem has been extensively studied in language understanding, however, it is still understudied in language generation, especially for pre-trained Seq2Seq models (Li et al., 2021). The popular adversarial learning methods for language understanding, such as FreeLB (Zhu et al., 2020), are not effective on Seq2Seq generation tasks (as shown in Table 4). Although few previous work have attempted to craft adversarial

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

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7160–7176
December 7-11, 2022 ©2022 Association for Computational Linguistics

samples for Seq2Seqs on machine translation (Belinkov and Bisk, 2018; Cheng et al., 2020), they have not been extensively studied across various generation tasks. Furthermore, they have not been studied on current pre-trained Seq2Seq models.

To address this problem, we provide the first quantitative analysis on the robustness of pre-trained Seq2Seqs. Specifically, we quantitatively analyze the robustness of BART across three generation tasks, i.e., text summarization, table-to-text, and dialogue generation. Through slightly modifying the input content, we find that the corresponding output significantly drops in both informativeness and faithfulness, which also demonstrate the close connection between the robustness of Seq2Seq models and their informativeness and faithfulness on generation tasks.

Based on the analysis above, we further propose a novel **Adv**ersarial augmentation framework for **Seq**uence-to-Sequence generation (AdvSeq) to enhance its robustness against perturbations and thus obtain an informative and faithful text generation model. AdvSeq constructs challenging and factually consistent adversarial samples and learns to defend against their attacks. To increase the diversity of the adversarial samples, AdvSeq applies two types of perturbation strategies, implicit adversarial samples (AdvGrad) and explicit token swapping (AdvSwap), efficiently utilizing the back-propagate gradient during training. AdvGrad directly perturbs word representations with gradient vectors, while AdvSwap utilizes gradient directions for searching for token replacements. To alleviate the vulnerability of NLL, AdvSeq adopts a KL-divergence-based loss function to train with those adversarial augmentation samples, which promotes higher invariance in the word representation space (Miyato et al., 2019).

We evaluate AdvSeq by extensive experiments on three generation tasks: text summarization, table-to-text, and dialogue generation. Our experiments demonstrate that AdvSeq can effectively improve Seq2Seq robustness against adversarial samples, which result in better informativeness and faithfulness on various text generation tasks. Comparing to existing adversarial training methods for language understanding and data augmentation methods for Seq2Seqs, AdvSeq can more effectively improve both the informativeness and faithfulness for text generation tasks.

We summarize our contributions as follows:

- To the best of our knowledge, we are the first to conduct quantitative analysis on the robustness of pre-trained Seq2Seq models, which reveal its close connection with their informativeness and faithfulness on generation tasks.

- We propose a novel adversarial argumentation framework for Seq2Seq models, namely AdvSeq, which effectively improves their informativeness and faithfulness on various generation tasks via enhancing their robustness.

- Automatic and human evaluations on three popular text generation tasks validate that AdvSeq significantly outperforms several strong baselines in both informativeness and faithfulness.

2 Seq2Seq Robustness Analysis

In this section, we analyze the robustness of the Seq2Seq by evaluating its performance on adversarial samples. In brief, after the input contexts are minorly modified, we check whether the model maintains its informativeness and faithfulness. A robust model should adaptively generate high-quality texts corresponding to the modified inputs.

Following the definition of adversarial examples on Seq2Seq models, *adversarial examples should be meaning-preserving on the source side, but meaning-destroying on the target side* (Michel et al., 2019). Formally, given an input context x and its reference text y_{ref} from the test set of a task, and a Seq2Seq model f_θ trained on the training set, we fist collect the original generated text $y = f_\theta(x)$. We measure its faithfulness and informativeness by $E_f(x, y)$ and $E_i(x, y, y_{ref})$, where E_f and E_i are the faithfulness and informativeness metrics, respectively. Then, we craft an adversarial sample x' by slightly modifying x trying not to alter its original meaning and generate y' grounded on x'. Finally, we measure the *target relative score decrease* (Michel et al., 2019) of faithfulness after attacks by:

$$d = \frac{E_f(x, y) - E_f(x', y')}{E_f(x, y)} \quad (1)$$

We calculate the decrease of informativeness similarly. We also report the entailment score of x' towards x: $S(x, x')$ to check whether the modification changes the meaning.

Task	S	E_i	E_f
Summarization	EntS	ROURGE-L	CC/QE
Table-to-text	-	PARENT	PARENT
Dialogue NLI	EntS	PPL	PPL/Ranking

Table 2: Evaluation metrics for different tasks, where we apply entailment score (EntS) for S, FactCC (CC), QuestEval (QE) for E_f of summarization and ROUGE-L (R-L) for E_i of summarization. PPL is the abbreviation for perplexity.

Evaluation Settings We apply BART as f_θ and conduct evaluations on three datasets, XSum for text summarization, WIKIPERSON for table-to-text, and Dialogue NLI for dialogue generation, with 2,000 samples sampled from each dataset. Evaluation metrics for different tasks are listed in Table 2, and the details are introduced in §4.

As a preliminary study of robustness, we design a simple word swapping-based method for crafting adversarial samples. For word $w_i \in x$, we first calculate its salience score as $f_\theta(y_{ref}|x) - f_\theta(y_{ref}|x \setminus w_i)$, where $f_\theta(y|x)$ is the validation score of generating y given x, and $x \setminus w_i$ is input x with word w_i deleted. We then sort the salient score to get the swapping orders of words in x, giving priority to those with higher scores. Following the orders, we iteratively swap each word with its top 10 nearest neighbors in the word embedding space and keep the replacement if the output BLEU score decreases after swapping. For better meaning preservation, we hold the maximum difference in edit distance to a constant of 30 for each sample. The details of algorithm and evaluation metrics are introduced in Appendix A.

Attack Results Reported in Table 3, for informativeness, text summarization drops by 8.75% in ROUGE-L, dialogue generation increases by 16.48% on the reference perplexity, table-to-text drops by 4.41% and 6.11% on the PATENT recall and F-1, respectively. For faithfulness, text summarization drops by 15.53% and 5.67% in FactCC and QuestEval, table-to-text generation drops in PARENT precision by 6.24%, dialogue generation increase in the perplexity of entailed candidates by 5.67%. Overall, BART is still not robust enough to tackle minor perturbations in the input, which lead to degeneration of the generalization performance of both informativeness and faithfulness.

Input	PARENT		
	Precision	Recall	F-1
Ori.	57.61	97.54	71.56
Adv.	54.01	93.17	67.19
d	$6.24^{\dagger}$	$4.41^{\dagger}$	$6.11^{\dagger}$

(a) Table-to-text (WIKIPERSON)

Input	Perplexity			Hit@1↑
	Ref↓	Ent.↓	Con.↑	
Ori.	10.30	22.08	16.15	37.64
Adv.	12.01	22.75	16.60	35.62
d	$-16.48^{\dagger}$	$-2.95^{\dagger}$	0.25	$5.67^{\dagger}$

(b) Dialogue Generation (Dialogue NLI)

	EntS	R-L	CC.	QE
Ori.	-	35.96	16.74	43.46
Adv.	83.6	32.83	14.18	41.20
d	-	$8.75^{\dagger}$	$15.53^{\dagger}$	$5.20^{\dagger}$

(c) Text Summarization (XSum)

Table 3: Evaluation performance of fine-tuned BART on original samples (Ori.) vs adversarial samples (Adv.), d is the target relative score decrease of every metric. †: significantly decrease (p < 0.01) by t-test.

3 AdvSeq Framework

Inspired by the findings in the previous section, we propose to improve the informativeness and faithfulness of a Seq2Seq model via enhancing its robustness. We propose our framework AdvSeq for robustly fine-tuning pre-trained Seq2Seq models. During training, AdvSeq utilizes gradient information to automatically construct challenging but factually consistent augmentation samples. For better diversity, we construct two kinds of augmentation samples, implicit adversarial sample (AdvGrad) and explicit token replacement (AdvSwap). In the following, we introduce AdvSeq in detail.

Given an input sample (x, y), a Seq2Seq model with parameters θ learns to generate fluent text by training with NLL loss. Considering the vulnerability of NLL, we further measure and optimize probability distribution changes caused by small random perturbations via KL divergence. The overall loss function w.r.t. the clean sample (x, y) is then defined as:

$$\mathcal{L}_o(x+\delta_x, y+\delta_y, \theta) = -\sum_{y_t \in y} \log p(y_t|x, y_{<t}) \tag{2}$$

$$+ KL_S(p(y_t|y_{<t}+\delta_y, x+\delta_x), p(y_t|y_{<t}, x))$$

where the first term is NLL, KL_S [1] is the symme-

[1] $KL_S(x, y) = KL(x|y) + KL(y|x)$

try of KL divergence, and perturbations δ_x, δ_y are sampled from a uniform distribution and added on word embeddings as default. *After that, we back-propagate $\mathcal{L}_o$ and apply its gradient to construct AdvGrad and AdvSwap.*

AdvGrad Directly utilizing back-propagated gradient of $\mathcal{L}_o$ as perturbations, we construct implicit adversarial samples. Instead of randomly perturbing word representations like $\mathcal{L}_o$, AdvGrad further searches for stronger perturbations that mostly affect the generation process by solving:

$$\min_{\theta} \mathbb{E}_{(x,y)}[\max_{\delta_x,\delta_y} \mathcal{L}_o(x+\delta_x, y+\delta_y, \theta)] \quad (3)$$
$$s.t. ||\delta_x||_F \leq \epsilon, \; ||\delta_y||_F \leq \epsilon$$

where we constrain the Frobenius norm of a sequence by ϵ. Because it is intractable to solve this equation, we perform gradient ascent (Madry et al., 2018) on the δ_x to approximate the solution:

$$\delta_x^{'} = \Pi_{||\delta_x||_F \leq \epsilon}(\delta_x + \alpha * g_x/||g_x||_F) \quad (4)$$
$$g_x = \nabla_{\delta_x} \mathcal{L}_o(x+\delta_x, y+\delta_y) \quad (5)$$

where δ'_x is the updated adversarial perturbation, α is the learning rate of the update, $\Pi_{||\delta_x||_F \leq \epsilon}$ performs a projection onto the ϵ-ball. $\delta_y^{'}$ is approximated though similar steps. Though previous works apply multi-step updates for a closer approximation (Madry et al., 2018; Zhu et al., 2020), we find one-step update is effective and efficient enough for AdvGrad. By perturbing the word embeddings with the adversarial perturbations: $x' = x + \delta_x^{'}$, $y' = y + \delta_y^{'}$, we get a pair of parallel augmentation samples, (x', y'). Because the perturbations are implicit and minor (within a ϵ-ball), we consider (x', y') to be meaning preserving. We then train with (x', y') by optimizing the KL divergence between its output word distributions $p(y'|x')$ with the original $p(y|x)$ by:

$$\mathcal{L}_i = \sum_{y_t \in y} KL_S(p(y_t|y'_{<t}, x'), p(y_t|y_{<t}, x)) \quad (6)$$

AdvSwap Simultaneously utilizing gradient directions of $\mathcal{L}_o$, we construct explicit adversarial samples by token swapping. The core idea is to identify and swap salient tokens in x without or minorly changing the original meaning of x. The first procedure is to identify a set of salient tokens in x that attributed most to the generation. We formulate this procedure as a causal inference problem (Yao et al., 2021; Xie et al., 2021b). Concretely, given target context $\mathcal{Y}$, which can either be the whole reference text y or sampled spans from it, *we need to infer a set of most relevant tokens $\mathcal{X}$ from x that contribute most for generating $\mathcal{Y}$*. Because the difficulties of this procedure depend on specific tasks, we apply two strategies for searching $\mathcal{X}$ in different tasks:

- **Gradient Ranking**: Select tokens $\{x_t \in x\}$ with highest $k\%$ two-norm of gradient $||\nabla_{x_i} \mathcal{L}_o(x, \mathcal{Y})||$.
- **Word Overlapping**: Find tokens in x that overlap with $\mathcal{Y}$: $x \cap \mathcal{Y}$.

For tasks like table-to-text that word overlapping is enough to infer the precise causality, we randomly sample several spans from y as $\mathcal{Y}$ and find the corresponding $\mathcal{X}$ by word overlapping. While, for highly abstractive generation like text summarization, we use the global information y as $\mathcal{Y}$ and infer $\mathcal{X}$ by gradient ranking, which measures the salience of a token by gradient norm (Simonyan et al., 2014; Li et al., 2016).

After the $\mathcal{X}$ is inferred, we search around the neighbors of word embedding space for meaning preserving swapping, utilizing the semantic textual similarity of word embeddings (Li et al., 2020a). To make the adversarial samples more challenging, we search at the direction of gradient ascent to replace $x_t \in \mathcal{X}$ with $\hat{x_t}$:

$$\hat{x}_t = \operatorname*{argmax}_{x_i \neq x_t} \cos(e_{x_i} - e_{x_t}, \nabla_{x_i} \mathcal{L}_o(x, \mathcal{Y})) \quad (7)$$

where e_{x_i} is the word embedding of x_i in the vocabulary list, and $\cos(\cdot, \cdot)$ is the cosine similarity of two vectors. After word swapping is done for all the tokens in $\mathcal{X}$, we get the adversarial sample x''. We train the explicit adversarial sample $(x'', \mathcal{Y})$ with KL divergence:

$$\mathcal{L}_e = \sum_{y_t \in \mathcal{Y}} KL_S(p(y_t|y_{<t}, x''), p(y_t|y_{<t}, x)) \quad (8)$$

Overall Training For efficiently utilizing the first back-propagate step of $\mathcal{L}_o$, we apply the "Free" Large-Batch Adversarial Training (FreeLB) (Zhu et al., 2020). For every training step, we first forward and back-propagate $\mathcal{L}_o$ for constructing AdvGrad and AdvSwap while saving the gradient $\nabla_\theta \mathcal{L}_o$ with respect to θ. Next, we forward and back-propagate the loss function of two augmentations: $\mathcal{L}_i + \mathcal{L}_e$ and accumulate its gradient with respect to model parameters θ. Finally, we update θ with previous two-step accumulated gradient. The overall training procedure is summarized in the Algorithm 2 in the appendix.

4 Experiment Setup

4.1 Datasets

We conduct experiments on four datasets of three text generation tasks: text summarization, table-to-text generation, and dialogue generation. For text summarization, we use XSum (Nallapati et al., 2016) and CNN/DM (Hermann et al., 2015) for evaluation. For table-to-text generation, we use WIKIPERSON (Wang et al., 2018). For dialogue generation, we apply dialogue NLI (Welleck et al., 2019). Details of all datasets are listed in Table 10.

4.2 Automatic Metric

Informative Metric We apply ROUGE F_1 (Lin, 2004) to evaluate text summarization, BLEU (Papineni et al., 2002) for table-to-text. For dialogue generation, given an example (x, y) consists of a dialogue history $x = \{p_1, \ldots, p_k, u_1, \ldots, u_t\}$ and reference utterance y, where p_i is a given persona sentence and u_i a dialogue utterance, we report the perplexity of generating reference response y to evaluate informativeness.

Faithfulness Metric For text summarization, since previous works (Pagnoni et al., 2021) report the unreliability of existing factual metrics, we report two different types of metrics, FactCC (CC) (Kryscinski et al., 2020) based on textual entailment and QuestEval (QE) (Scialom et al., 2021) based on question answering, for reliable comparison. For table-to-text generation, we report the PARENT (Dhingra et al., 2019) score of generated texts, which is a hybrid measurement of its faithfulness and informativeness. For dialogue generation, we report the perplexity of generating entailed and contradicted candidate utterances provided in the dialogue NLI. We also report **Hit@1** of ranking candidate utterances by their modeling perplexity (the lower the better), which is the probability that the top one candidate is entailed by the dialogue history. Each example in dialogue NLI contains 10 entailment and 10 conflict utterances.

4.3 Baseline Methods

In all of our experiments, we employ the pre-trained Seq2seq model BART (Lewis et al., 2020) as our base model, which is then fine-tuned with various augmentation methods for particular tasks

Adversarial Training While adversarial training has achieved promising results in natural language understanding, its performance on text generation lacks extensive evaluations. We evaluate two representative methods, **FreeAT** (Shafahi et al., 2019) and **FreeLB** (Zhu et al., 2020) on Seq2Seq tasks.

Faithfulness Augmentation **CLIFF** (Cao and Wang, 2021) is a contrastive learning-based method which learns to discriminate various heuristically constructed augmentation samples. Loss Truncation (**LT**) (Kang and Hashimoto, 2020) is a faithfulness augmentation method which adaptively removes noisy examples during training. **Aug-plan** (Liu et al., 2021a) augments table-to-text training by incorporating auxiliary entity information.

Common Augmentation We also compare AdvSeq with other common data augmentation methods for text generation. **R3F** (Aghajanyan et al., 2020) fine-tunes pre-trained language models with random noises on embedding representations. **SSTIA** (Xie et al., 2021a) constructs the target-side augmentation sample by a mix-up strategy.

5 Implementation Details

During fine-tuning, we set the learning rate to 3e-5, label smoothing to 0.1. For XSum and CNN/DM, we follow the parameter settings of Lewis et al. (2020). For WIKIPERSON and Dialogue NLI we apply the similar parameter settings with Lewis et al. (2020) on CNN/DM, except we do not apply trigram block during inference. For AdvGrad we set α to 4e-1, ϵ to 2e-1. For AdvSwap we randomly select two spans that contain overlapping tokens with x as $\mathcal{Y}$ for table-to-text, we set k to 0.15 for gradient ranking for text summarization and dialogue NLI.

6 Results

6.1 Automatic Evaluation

Text Summarization Table 4 reports results on XSum and CNN/DM. As demonstrated by the results of FreeLB and FreeAT, directly fine-tuning with adversarial samples by NLL loss does not benefit BART. Though they construct adversarial samples similarly based on gradient, AdvGrad achieves much better performance. Compared with all baselines in respective of informativeness, our mod-

Datasets	XSum					CNN/DM				
	R-1	R-2	R-L	CC	QE	R-1	R-2	R-L	CC	QE
BART	45.20	21.90	36.88	23.64	45.89	44.08	20.92	40.79	76.09	51.15
FreeAT	45.51	22.08	37.16	23.23	45.16	44.02	20.88	40.75	**77.69**	51.09
FreeLB	45.25	21.98	36.89	23.13	45.90	44.16	21.05	40.83	76.53	51.18
LT	44.42	21.10	36.17	23.99	45.29	44.05	20.89	40.79	77.06	50.97
CLIFF	44.63	21.39	36.43	23.51	45.45	44.29	21.14	41.02	75.66	50.47
R3F	45.47	22.02	37.24	**26.10**	45.83	44.31	21.14	41.00	76.64	51.14
SSTIA	45.82	22.42	37.50	23.89	46.05	44.76	21.46	41.57	71.69	51.20
AdvGrad	**46.13**	**22.72**	**37.81**	23.49	**47.82**†	44.57	21.50	41.36	75.06	50.67
AdvSwap	45.49	22.21	37.21	24.70	47.25†	44.63	21.47	41.35	74.46	51.21
AdvSeq	46.06	22.65	37.69	24.67	47.61†	**44.96**	**21.73**	**41.79**	72.43	**51.50**†

Table 4: Experimental results on text summarization, where R-1 (ROUGE-1), R-2 (ROUGE-2), R-L (ROUGE-L) are informative metrics and CC, QE report faithfulness. The last block reports the results of our methods. The underlines indicate the second-best performance. †: significantly better than all the baseline model ($p < 0.01$).

	BLEU	**PARENT**
Wang et al. (2020)	24.56	53.06
BART	**31.30**	56.40
LT	29.70	56.69
Aug-plan	17.12	56.75
R3F	31.08	56.56
SSTIA	30.34	56.84
AdvGrad	31.17	56.81
AdvSwap	29.95	56.94
AdvSeq	30.37	**57.33**†

Table 5: Experimental results on WIKIPERSON. Note that the reference text in this dataset may contain hallucinated content, so BLEU scores can not measure the faithfulness of generation.

els produce the best ROUGE scores. Compared with all baselines in respective of faithfulness, our models also produce significantly better QE scores, which correlate better with human judgments than CC scores on both datasets, as claimed by Cao and Wang (2021). Specifically, both AdvGrad and AdvSwap improve ROUGE and QE scores across datasets. Through combing them, AdvSeq produces a better performance on CNN/DM.

Table-to-text Due to the existence of noisy training samples, PARENT is considered a better metric than BLEU by matching the content with both the input table and reference text (Dhingra et al., 2019). As reported in Table 5, AdvSeq obtains significant higher PARENT scores comparing with all baselines. Both AdvGrad and AdvSwap improve PARENT scores, and through combining them, AdvSeq further achieves the best performance.

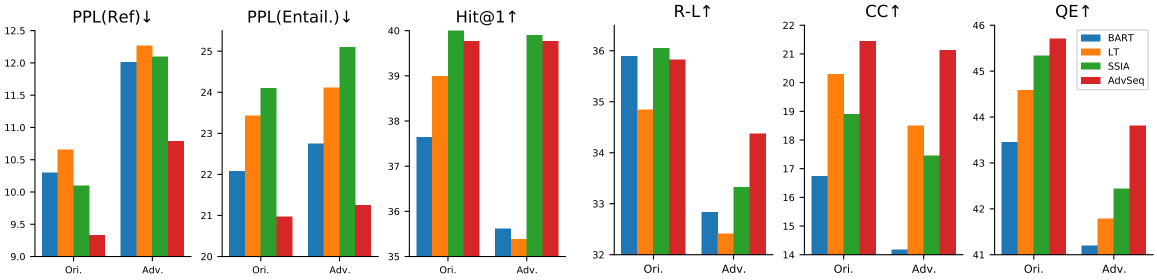
	Perplexity			Hit@1
	Ref.↓	Ent.↓	Con.↑	Ent.%↑
BART	10.9	22.2	16.1	36.5
LT	11.2	23.6	16.7	38.0
SSTIA	10.7	24.3	**18.4**	**40.9**
AdvSeq	**9.8**†	**21.0**†	15.9	38.7

Table 6: Experimental results on Dialogue NLI, where Ref, Ent and Con indicates perplexity on reference, entailed and contradict candidate utterances, respectively.

Dialogue Generation As reported in Table 6, AdvSeq produces the lowest perplexity on reference and entailed candidates, exceeding the second ranked systems by 8.4% and 5.4%, respectively. As AdvSeq mainly improves faithfulness by learning from factually consistent samples, it does not help with distinguishing contradictory samples. Thus, AdvSeq improves Hit@1 over BART and surpasses LT by producing lower perplexity on entailed candidates without changing the perplexity of contracted candidates. Overall, the results indicates that AdvSeq improves both informativeness and faithfulness of dialogue generation.

6.2 Human Evaluation

We recruit two human annotators from a professional data annotation team to evaluate our generated text in both informativeness and faithfulness. The evaluations are conducted on text summarization and table-to-text generation. All the annotators are first trained with the given criteria and demonstrations. After they are tested to be familiar with these criteria, we randomly select generated samples for annotation. For each sample, the annotators



(a) Dialogue generation (Dialogue NLI)

(b) Text Summarization (XSum)

Figure 1: Robust Analysis over four systems on dialogue generation and text summarization.

	Faithful.			Inform.		
Model	Win↑	Tie	Lose↓	Win↑	Tie	Lose↓
SSTIA	20.0	63.5	**16.5**	23.5	52.0	24.5
CLIFF	16.0	59.0	25.0	25.5	52.0	**22.5**
AdvSeq	**27.0**	53.5	19.5	**33.5**	43.5	23.0

(a) XSum

	Faithful.			Inform.		
Model	Win↑	Tie	Lose↓	Win↑	Tie	Lose↓
SSTIA	25.0	57.5	**17.5**	21.0	64.0	**15.0**
Aug-plan	10.5	22.0	67.5	18.0	27.0	55.0
AdvSeq	**40.0**	41.0	19.0	**31.5**	52.5	16.0

(b) WIKIPERSON

Table 7: Percentage of generated text that are better than, tied with or worse than BART, in faithfulness (Faithful.) and informativeness (Inform.). The Cohen's kappa scores are 63.46 and 60.35 for two aspects on XSum, and 52.89 and 49.47 for WIKIPERSON.

are shown generated texts from BART and other three systems, a data augmentation method (SSTIA), a faithfulness improvement method (CLIFF or Aug-plan), and our AdvSeq. The annotators are asked to judge whether these generated texts are better than, tie with, or worse than the baseline model from the informativeness and faithfulness aspects, separately. Evaluation results of 100 randomly selected samples from each dataset are reported in Table 7. Compared with other baselines, *model trained with AdvSeq is more frequently rated as being more informative and more faithful.*

7 Analysis

7.1 Robustness Analysis

As demonstrated in Figure 1, when testing on adversarial samples, models trained with AdvSeq degenerate the least and achieve the best results in almost all the metrics across tasks. Especially for dialogue generation, Hit@1 performance does not decrease on adversarial samples for AdvSeq. In summary, robustness analysis on three tasks validates that AdvSeq better improves robustness compared with other baselines.

	R-1	R-2	R-L	CC	QE
AdvGrad	**46.13**	**22.72**	**37.81**	23.49	**47.82**
w/o δ_y	45.81	22.39	37.47	**24.56**	47.56
w/o δ_x	46.08	22.65	37.70	23.85	47.75
w/o KL_S	45.25	21.98	36.89	23.13	45.90
BART	45.20	21.90	36.88	23.64	45.89

Table 8: Ablation study for AdvGrad on XSum.

7.2 AdvGrad Analysis

Ablation Study We study how each component affects the performance of AdvGrad on XSum, reported in Table 8. We observe that, first, *perturbations on encoder and decoder both benefit AdvGrad.* After removing δ_x or δ_y, every metric score drops (except CC). Second, *KL is crucial for training with adversarial samples.* Removing KL degrades faithfulness and informativeness into the similar performance with BART (w/o KL vs BART).

Perturbed Layer We further study how perturbing representations from different layers affect generation performance. We conduct experiments on AdvGrad by gradually moving δ_x or δ_y from word embeddings to high-layer representations of the encoder or decoder, separately. In this procedure, the other side perturbation remains fixed on the word embedding. The results are illustrated in Figure 2, where the horizontal axis (Layer id) indicates the perturbed layer and the longitudinal axis reports the corresponding ROUGE-L score.[2] Because both BART encoder and decoder are composed of 12 transformer layers, we denote layer 0 as the word

[2] We test on even number layers.

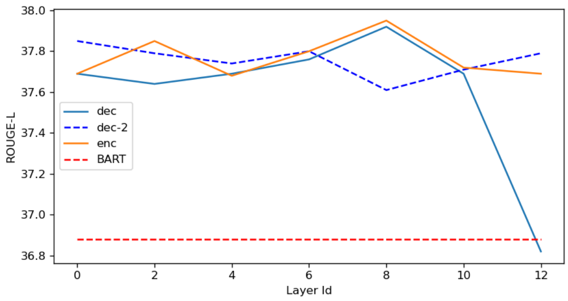


Figure 2: ROUGE-L scores of AdvGrad on XSum when training with perturbations on representations of different encoder (**enc**) or decoder (**dec**) layers. **dec-2** indicates training with two-step gradient ascent perturbations on the decoder .

$\mathcal{Y}$	$\mathcal{X}$	BLEU	**PARENT**
Spans	Word Overlapping	30.37	57.33
Spans	Gradient Ranking	29.68	56.84
y	Word Overlapping	31.43	56.59

(a) Table-to-text, WIKIPERSON

$\mathcal{Y}$	$\mathcal{X}$	ROUGE-L	FactCC
y	Gradient Ranking	37.81	24.67
y	Word Overlapping	37.37	24.17
Spans	Gradient Ranking	37.56	24.24

(b) Text summarization, XSum

Table 9: Analysis on different settings of AdvSwap, where we report how the performance of AdvSeq changes when changing $\mathcal{X}$ and $\mathcal{Y}$.

embeddings and layer 12 as the output representations of the encoder or decoder.

From the results, we conclude that *in both encoder and decoder, perturbing the middle-layer representations achieves the best results.* When layer id increases, ROUGE-L scores first increase to the peak at layer 8, and then decrease. In particular, when attacking on the output of the decoder (layer 12 of **dec**), ROUGE-L score significantly drops because the perturbation only affects the final word prediction layer.

Multi-step Attack We also study how multi-step gradient ascent (repeating Eq.5) affects generation. We perform two-step gradient ascent on perturbing the decoder. From Figure 2, we can observe that comparing to one step ascent (**dec**), two-step ascent (**dec-2**) further benefits lower layers (layer 0, 2, 4) but does not help or even worsen higher layers (layer 6, 8, 10). Overall, AdvGrad still has great potential. Careful selection of permuted layers and gradient ascent steps will further boost its performance.

7.3 AdvSwap Analysis

We analyse how different settings in AdvSwap affect the overall performance of AdvSeq. In Table 9, two factors are studied: selection of $\mathcal{Y}$, inference strategy for $\mathcal{X}$. From the results, we can draw into conclusion: *More precise casualty inference lead to better performance, which depends on both inference strategies ($\mathcal{X}$) and the selection of $\mathcal{Y}$.* In table-to-text, we can precisely infer $\mathcal{X}$ by word overlapping, since the correctly generated facts are mainly copied from the input. While for tasks like XSum, the summary is too abstract to infer precisely. Thus, changing the searching strategies (second rows in both Table 9a, 9b) lead to drops in all metrics. For selecting $\mathcal{Y}$, using the whole y with word overlapping in table-to-text will cause all the value tokens in table x be selected into $\mathcal{X}$, which lead to worse performance. While in summarization, using global information in y with gradient ranking produces the best performance.

8 Related Works

Faithfulness in Text Generation Recently, a variety of works in text generation tasks have realized the severity of the unfaithful generation problem. Improving the faithfulness of text generation has been an important and popular topic (Li et al., 2022). Similar strategies have been proposed for different tasks. For example, factual guidance is one of the most popular methods. Cao et al. (2018); Li et al. (2020c); Wu et al. (2021a) leverage guidance information like keywords and knowledge graph for faithful summarization; Rashkin et al. (2021); Wu et al. (2021b) respectively guide response generation by evidence sentence and control phrases. For data augmentation based methods, Liu et al. (2021b) utilize entity swapping, while Cao and Wang (2021) further construct various kinds of augmentation samples including entity regeneration, masking relation, etc.

Adversarial Learning Adversarial attack (Goodfellow et al., 2015) searches for imperceptible perturbations to mislead neural networks. Adversarial training (Goodfellow et al., 2015), derived from gradient-based adversarial attacks, trains the model with generated adversarial samples. Besides extensive exploration in computer vision (Carlini and Wagner, 2017), a large number of studies on adversarial learning for NLP emerge recently. While most of them focus on natural language understanding tasks (Ebrahimi et al., 2018; Jia and Liang, 2017; Li et al., 2020b), some works also craft adversarial samples for Seq2Seq models (Michel et al., 2019; Cheng et al., 2020). For adversarial training, although several recent works apply gradient-based methods to fine-tune pre-trained NLU models (Zhu et al., 2020; Jiang et al., 2020), adversarial analysis and learning for pre-trained Seq2Seq models is still lack of exploration.

Data Augmentation Data augmentation has been widely applied in almost every subarea in NLP. Several studies propose universal data augmentation methods for text generation tasks. Back-translation (Sennrich et al., 2016) constructs augmentation samples via auxiliary translation models, which are widely used in improving the quality of machine translation. R3F (Aghajanyan et al., 2020) inject random noise on the word embeddings of the pre-trained language model for robust fine-tuning. SSTIA (Xie et al., 2021a) applies a mix-up strategy on the target side of Seq2Seq to construct augmentation examples. These methods mainly focus on improving the informativeness of generated text and lack of faithfulness analysis.

9 Conclusion

We are the first to conduct quantitative analysis on the robustness of pre-trained Seq2Seq models, and reveal the close connection between the robustness of Seq2Seq models and the informativeness and faithfulness for text generation tasks. Through quantitative analysis, we point out that current SOTA pre-trained Seq2Seq model is still not robust enough to tackle minor perturbations in the input. To alleviate this problem, we propose AdvSeq to improve the faithfulness and informativeness of Seq2Seq models via enhancing their robustness. Augmented by two types of adversarial samples during training, AdvSeq generates text with much better informativeness and faithfulness compared to strong baselines on text summarization, table-to-text, and dialogue generation.

Limitations

AdvGrad In AdvGrad, we only separately approximate the optimal perturbations for x and y with first order gradient. By exploring relations between perturbations on encoder and decoder sides, more efficient and effective approximate solutions of Eq.4 may be given.

Computational Cost Compared with standard fine-tuning, AdvSeq constructs adversarial samples with one extra backward pass and three extra forward passes. AdvSeq requires approximately 4 times computation time and 1.7 times GPU memory usage for one batch than standard fine-tuning on BART.

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 (2019YFB1704002) and National Natural Science Foundation of China (61876009).

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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.

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.

Bhuwan Dhingra, Manaal Faruqui, Ankur P. Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. 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 4884–4895. 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.

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*.

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.

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.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: robust and efficient fine-tuning for pretrained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2177–2190. 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, ACL 2020, Online, July 5-10, 2020*, pages 718–731. 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.

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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 681–691. The Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2105.10311*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. 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.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *CoRR*, abs/2203.05227.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020c. Leveraging graph to improve abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6232–6243, Online. Association for Computational Linguistics.

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

Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021a. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13415–13423. AAAI Press.

Wei Liu, Huanqin Wu, Wenjing Mu, Zhen Li, Tao Chen, and Dan Nie. 2021b. Co2sum: Contrastive learning for factual-consistent abstractive summarization. *CoRR*, abs/2112.01147.

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 T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. 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 3103–3114. Association for Computational Linguistics.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.

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.

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

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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 704–718. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Qingyun Wang, Xiaoman Pan, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji, and Kevin Knight.

2018. Describing a knowledge base. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 10–21, Tilburg University, The Netherlands. Association for Computational Linguistics.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086, Online. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Ziqiang Cao, Sujian Li, Hua Wu, and Haifeng Wang. 2021a. BASS: Boosting abstractive summarization with unified semantic graph. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6052–6067, Online. Association for Computational Linguistics.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and Bill Dolan. 2021b. A controllable model of grounded response generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14085–14093. AAAI Press.

Shufang Xie, Ang Lv, Yingce Xia, Lijun Wu, Tao Qin, Tie-Yan Liu, and Rui Yan. 2021a. Target-side input augmentation for sequence to sequence generation. In *International Conference on Learning Representations*.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021b. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 100–110. Association for Computational Linguistics.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Trans. Knowl. Discov. Data*, 15(5):74:1–74:46.

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

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Analysis Details

Evaluation Metrics For text summarization, we apply a Roberta classifier fine-tuned on MNLI dataset as $S_f(x, x')$ for predicting whether the modified sentences in x' are factually consistent with their originals in x.

For table-to-text, because its inputs are tables of facts, it is hard to modify them without changing their meaning, we only report how adversarial examples affect faithfulness by PARENT. PARENT-precision measures how much generated contents are covered by the input table or the reference text. PARENT-recall measures how much contents in the input table or the reference text are mentioned in the generated context.

Adversarial Sample Details of crafting adversarial samples are introduce in Algorithm 1.

Algorithm 1: Adversarial Sample for Seq2Seq

```
Require : Original Sample $(x, y_{ref})$,
          Fine-tuned model $f_\theta$
$y = f_\theta(x)$
$current\_bleu = BLEU(y, y_{ref})$
if $current\_bleu < 0.5$ then
    break
                                  ▷ Compute salient scores
for $w_i$ in $x$ do
    $s_i = f_\theta(y_{ref}|x) - f_\theta(y_{ref}|x \setminus w_i)$
$attack\_order = sorted(s)$
$x' \leftarrow x$
                                  ▷ Start attack
for $i$ in $attack\_order$ do
    if $current\_bleu = 0$ or
      $Edit\_distance(x', x) > 30$ then
        return $x'$
                    ▷ Search in top 10 nearest neighbors
    for $\hat{w}_i$ in $Nearest\_Neighbor(w_i, 10)$ do
        $x'' \leftarrow replace(x', \hat{w}_i)$
        $y' = f_\theta(x'')$
        $s' = BLEU(y', y_{ref})$
                                  ▷ End condition
        if $s' < current\_bleu$ then
            $current\_bleu \leftarrow s'$
            $x' \leftarrow x''$
            $break$
return $x'$
```

B AdvSeq

Details of AdvSeq are illustrated in Algorithm 2. Some cases of augmentation samples constructed by AdvSwap are demonstrated in Table 11. From these samples we can see, AdvSwap mainly swaps subwords with their lexically or semantically similar candidates, which not only bring about diversity but also preserve the original meaning.

Algorithm 2: AdvSeq

```
Require : Input sample $(x, y)$, model with
          parameters $\theta$
▷ Objective function for the original sample
$\delta_x, \delta_y \leftarrow U(\text{-1e-2,+1e-2})$
$\mathcal{L}_o \leftarrow$ Eq.3 with $(x, y, \delta_x, \delta_y)$
  ▷ Back-propagate $\mathcal{L}_o$ and save gradient of
  parameters $\theta$
$g_0 \leftarrow \frac{1}{2}\nabla_\theta \mathcal{L}_o(x, y)$
                                        ▷ AdvGrad
$\delta'_x \leftarrow \Pi_{||\delta_x||_F \le \epsilon}(\delta_x + \alpha * g(\delta_x)/||g(\delta_x)||_F)$
$\delta'_y \leftarrow \Pi_{||\delta_y||_F \le \epsilon}(\delta_y + \alpha * g(\delta_y)/||g(\delta_y)||_F)$
$\mathcal{L}_i \leftarrow$ Eq.6 with $(x + \delta'_x, y + \delta'_y)$
                                ▷ Construct AdvSwap
Set $\mathcal{Y}, \mathcal{X}$ depend on the specific task
Swap $\mathcal{X}$ in $x$ on Eq.7 to construct $x''$
$\mathcal{L}_e \leftarrow$ Eq.8 with $(x'', \mathcal{Y})$
                                   ▷ FreeLB training
Back-propagate $\mathcal{L}_i + \mathcal{L}_e$ and accumulate
  gradient of parameters $\theta$:
$g_1 \leftarrow g_0 + \frac{1}{2}\nabla_\theta(\mathcal{L}_i + \mathcal{L}_e)$
             ▷ Update the $\theta$ with learning rate $\tau$
$\theta \leftarrow \tau g_1$
```

C Datasets

The statistic details of all the datasets for experiments all reported in Table 10.

D Case Study

We demonstrate one case from XSum in Table 12 and two other cases in Table 13 and 14.

Datasets	#Train	#Valid	#Test
XSum	204,045	11,332	11,334
CNN/DM	287,227	13,368	11,490
WIKIPERSON	250,186	30,487	29,982
Dialogue NLI	10,013	968	12,376

Table 10: Statistics of datasets for evaluation.

	AdvSwap Cases (XSum)
Original x:	The accident happened in 2009 during a play dress rehearsal at the County Londonderry school . The family had sued the Western Education and Library Board for alleged negligence. A judge ruled that the school had appropriately supervised the rehearsal . The judge told the court that he backed the family's account that the child had lost sight in his left eye because he was struck by a girl holding a wand , who was said to have been casting a spell at the time. The boy , who can not be named for legal reasons , had been onstage with almost 200 other pupils getting ready for a performance . However , the judge held that teachers had properly assessed the wand and dismissed the claim . He said : " I do not consider the plaintiff has established any fault on the part of the defendant . "The judge praised the boy and his mother for the honesty of their evidence , but added there was " overwhelming evidence " that the girl who was holding the fairy wand was " timid and not likely to behave in an inappropriate way " .
AdvSwap x''	The accidents happened in 2009 during a play dress rehearsal at The Londond school . The family had sued theWestern Education and Library Board for alleged negligence. A judge ruled that the school had appropriately supervised the rehearsal . The judge told the court that he backed the family's account that the child had lost sight in his left eye because he was strike girls holding a Wand , who was said to have been casting a spell atthe time . The boy , who can not been named for legal reasons , had been with almost 200 other pupils getting ready for a performance . However , the judge held that teachers had properly assessed the wand and dismissed the claim . He said : " I do not consider the plaintiff has established any faultsOf The defendant . "The judge praised the boy and hismother for the honesty of their evidence , but added there was " overwhelming evidence " that the girl who was holding the Fairy wand was " timid and not likely behaves InAn Way " .
	AdvSwap Cases (WIKIPERSON)
Original x:	< Name ID > Kevin Wheatley < award received > Victoria Cross < date of birth > 13 March 1938 < date of death > 13 November 1965 < place of birth > Surry Hills, New South Wales < military branch > Australian Army
AdvSwap x''	<NameID>Kevin wheatleys< award received > Victoria Cross < date of Birth> 13 March 1938 < date of death> 13 November 1965 < place of birth >SurRY hills, New South Welsh < military branch > Australian Army
Original x:	<Name ID >Albert Bierstadt < country of citizenship > Germany < country of citizenship > United States < date of birth > January 7 1830 < date of death > February 18 1902 < genre > Landscape painting < conflict > American Civil War
AdvSwap x''	<Name ID >Albert Bierstadt < country of citizenship > Germany < country of citizenship > United States< dates of Birth >January7 1830< dates of death>February18 1902 < genre > Land Painting < conflict> American Civil war

Table 11: Cases from AdvSwap

Adversarial Attack Cases (XSUM)	
Original x:	The Hangzhou Internet Court opened on Friday and heard its first case - a copyright infringement dispute between an online writer and a web company.Legal agents in Hangzhou and Beijing accessed the court via their computers and the trial lasted 20 minutes. ... In 2016, China began streaming some trials in more traditional courtrooms online in an apparent effort to boost the transparency of the legal system. ... .In some other countries, online portals to allow people to resolve legal disputes in cyber-space already exist.Canada's Civil Resolution Tribunal starting accepting claims for 5,000 (Â£3,000) or less in British Columbia in June
BART Output	China's first internet court has opened its doors to the public for the first time in the world's largest city, state media report. .
AdvSeq Output	China has become the first country in the world to set up a court that allows people to sue each other via the internet , rather than in a traditional courtroom .
Adversarial x':	The Internet Court opened on Friday and heard its first case - a copyright infringement dispute between an online writer and a web company.Legal agents in Hangzhou and Beijing accessed the court via their computers and the experiment lasted 20 minutes....."In 2016, China began streaming some trials in more traditional courtrooms online in an apparent effort to boost the transparency of the legal system....In some other countries, online portals to allow public to resolve legal disputes in cyber-space already exist.Canada's Civil Resolution Tribunal starting accepting claims for 5,000 (Â£3,000) or less in British Columbia in June.
BART Output	A court in South Korea has opened its doors to the public for the first time in the world's first internet habeas court, state media report..
AdvSeq Output	China has become the first country in the world to open an internet - based court , which allows the public to take part in court cases via the internet.

Table 12: A case from summarization

Adversarial Attack Cases (WIKIPERSON)	
Original x:	< Name ID > Adam Darski < date of birth > 10 June 1977 < place of birth > Gdynia < country of citizenship > Poland < genre > Black metal < member of > Behemoth (band)
BART Output:	Adam Darski (born 10 June 1977 in Gdynia Poland) is a Polish musician best known as the lead vocalist of the Black metal band Behemoth (band) . Adam Darski was born on 10 June 1977 in Gdynia Poland . Behemoth (band) was formed in the early 2000s by Darski 's father Jan Darski and his brother Jan Darski .
AdvSeq Output:	Adam Darski (born 10 June 1977 in Gdynia Poland) is a Polish musician best known as the lead vocalist of the Black metal band Behemoth (band) . Adam Darski was born in Gdynia Poland on 10 June 1977 . Behemoth (band) was formed in the early 1990s in Gdynia Poland.
Adversarial x':	< Name ID > Damien Darski < date of birth > 10 April 1977 < place of birth > Gdynia < country of citizenship > Poland < genre > Black metal < member of >Dream (band)
BART Output:	Damien Darski (born 10 April 1977 in Gdynia Poland) is a Polish musician best known as the lead vocalist of the Black metal band Dream (band) . Damien Darski was born on 10 April 1977 in Gdynia Poland . Dream (band) was formed in 1999 in Gdynia Poland. Dream (band) was formed in 1999 in Poland.
AdvSeq Output:	Damien Darski (born 10 April 1977 in Gdynia Poland) is a Polish musician best known as the lead vocalist of the Black metal band Dream (band) . Damien Darski was born on 10 April 1977 in Gdynia Poland . Dream (band) 's debut album Dream (band) was released in 2006 .

Table 13: A case from table-to-text

Adversarial Attack Cases (WIKIPERSON)	
Original x:	< Name ID > Achille Raffray < occupation > Diplomat < date of birth > 17 October 1844 < date of death > 25 September 1923 < country of citizenship > France < place of death > Rome < member of > Société entomologique de France
BART Output:	Achille Raffray (17 October 1844 – 25 September 1923) was a France entomologist and Diplomat . Achille Raffray was born in Paris on 17 October 1844 . Achille Raffray died in Rome on 25 September 1923 . he was a member of the Société entomologique de France and the Société de France .
AdvSeq Output:	Achille Raffray (17 October 1844 – 25 September 1923) was a France Diplomat and entomologist . he was a member of the Société entomologique de France. he was born in Paris and died in Rome .
Adversarial x':	< Name ID > Jacopo Raffray < occupation > Albright < date of birth > 17 June 1844 < date of death > 25 Ninth 1923 < country of citizenship > France < place of death > Rome < member of > Société entomologique de France
BART Output:	Jacopo Raffray (17 June 1844 – 25 Ninth 1923 Rome) was a France entomologist . he was a member of the Société entomologique de France and the Société entomologique de France . raffray was born in Rio de Janeiro Brazil on 17 June 1844 and died in Rome on 25 Ninth 1923 . he was a member of the Société entomologique de France and the Société astronomique de France . he studied at the École Normale Supérieure and the École pratique des hautes études en sciences sociales in Paris and became an Albright
AdvSeq Output:	Jacopo Raffray (17 Jane 1844 – 25 Ninth 1923 Rome) was a France entomologist and Albright . he was a member of the Société entomologique de France. he was born in Rio de Janeiro Brazil and died in Rome Italy.

Table 14: A case from table-to-text