# GOLD-FACTUAL: Learning to Generate Faithful Summaries from Models' Generations

**Liyan Tang**
lytang@utexas.edu

**Zifan Xu**
zfxu@utexas.edu

## Abstract

In this paper, we propose GOLD-FACTUAL, a new training framework for text summarization that addresses the hallucination problem in abstractive summarization. The framework is built upon GOLD (Pang and He, 2021), an existing *offline* reinforcement learning training framework that is originally designed to alleviate long-standing problems in conventional maximum likelihood estimation (MLE) training. Instead, GOLD-FACTUAL leverges the *offline* training of GOLD and human evaluation feedback of factuality annotations that significantly improves the factual consistency of generated summaries. More specifically, a non-factual penalty reward is designed based on the factuality annotations of sub-optimal summaries generated by pre-trained text summarization models. Training on these *offline* demonstration summaries, GOLD-FACTUAL can improve factual consistency in the generated summaries and even outperform human reference summaries [1].

**Article:**
A total of 387 people were arrested between February 2016 and February 2017 - up from 255 the previous year. Meanwhile more than half of cabin crew who responded to a survey said they had witnessed disruptive drunken passenger behaviour at UK airports. …

**BART Summary:**
The number of people arrested for drink-driving at UK airports has more than doubled in the past year, according to figures obtained by BBC Panorama.

**GOLD-FACTUAL Summary:**
The number of people arrested at UK airports on suspicion of drinking alcohol in duty-free shops has risen, according to police figures obtained by Panorama.

Figure 1: An example of summarization outputs from BART and GOLD-FACTUAL for a news article in XSum. Orange and red text highlights the support and inconsistency for the highlighted sentence in the source article.

## 1 Introduction

Reinforcement Learning (RL) algorithms are developed to solve sequential decision making problems, where an agent has to achieve the goal by making not only one decision, but a sequence of decisions that influence the outcomes of a long future trajectory. Similar problems that deal with sequential output have been widely researched in the literature of Natural Language Processing (NLP) such as conditional text generation. Provided by the sequential nature of these problems, some prior works have applied RL algorithms in NLP problems (Wu et al., 2018; Ziegler et al., 2019; Cao et al., 2021). In this project, we plan to investigate in depth under the context of conditional text generation, which by nature is a sequential decision making problem that requires a model to generate texts conditioned on some prefixes.

Current approaches to conditional text generation largely rely on autoregressive language models trained with maximum likelihood estimation (MLE). However, this paradigm leads to two common challenges: (1) MLE tends to over-generalize, assigning large probability mass to both high-quality as well as low-quality and noisy sequences (Cao et al., 2021; Huszár, 2015); (2) MLE causes exposure bias problem: during training, the autoregressive model conditions on the ground truth history/prefix; however, at inference time it conditions on model-generated history.

To alleviate the aforementioned two challenges, Pang and He (2021) formulated text generation as an *offline* RL problem with expert demonstrations

[1] The video recording is at https://youtu.be/y6BD0nXkAcc.

(i.e., the reference) and introduced GOLD, a new *offline* RL training framework, which is shown to address the two challenges by both qualitative analysis and empirical results.

Our work builds on Pang and He (2021) and the GOLD framework [2]. GOLD is an *offline* policy gradient algorithm, which does not necessarily use expert demonstrators as the behavior policy. Therefore, GOLD framework has the potential to directly learn a text generation policy from near-optimal human demonstrations as long as appropriate reward labels are given. In this work, we propose GOLD-FACTUAL that addresses the hallucination problem in abstractive summarization. Factual hallucination in the generated text is hard to avoid under the MLE training objective. Several works (Böhm et al., 2019; Stiennon et al., 2020) tackle this problem via learning a reward function to approximate human preference over summary pairs. Instead, GOLD-FACTUAL seeks to directly optimize the text summarization model from sub-optimal demonstration but with ground truth reward labels from human. Specifically, we leverage several datasets that have token-level human annotations of factuality errors and then train a model to generate less hallucinated text. Extra factuality labels on human reference summaries is shown to improve the performance of summarization models by learning a reward function that measures the factuality (Cao et al., 2021). In summary, GOLD-FACTUAL uses factuality labels from generated summaries and the GOLD framework to learn the text generation model in an *offline* fashion.

## 2 Related Work

Conventional NLP approaches train the text generation models to maximize the likelihood estimation of human reference data, which can be thought of as simple behavior cloning approaches that clone human text generation from a perspective of imitation learning. Instead, we focus on the approaches that explicitly solve the text generation problem with RL approaches. These approaches can either be *offline* training frameworks that still learn from fixed demonstration data but more explicitly tackle the sequential nature of the problem (Section 2.1), or *online* training frameworks that learn from models' own generations and compute rewards based on existing text generation metrics (Section 2.2). In addition, Section 2.3 introduces approaches with human evaluation feedback to encourage the model to generate texts that better matches human quality. In the end, Section 2.4 briefly introduces factual hallucinations, a well-known problem text generation that can be solved with the proposed approach.

### 2.1 *Offline* Training Framework

Pang and He (2021) proposed GOLD, an *offline* policy optimization framework, which treats the expert text generation as a behavior policy and performs *off-policy* policy gradient update with the trajectories from the behavior policy. Those trajectories are pre-collected and fixed without an explicit reward label, so the whole framework falls in the paradigm of imitation learning. To guarantee optimizing the optimal policy, the policy is updated with importance weighting under some necessary approximations. Even though the framework does not employ an external reward function, it designs a reward function $R(a_t, s_t) = p_{\text{MLE}}(a_t, s_t)$ based on $p_{\text{MLE}}$, the maximum likelihood probability pretrained supervisely that approximates a truth human probability $p_{\text{human}}$. Intuitively, this reward function not only encourages the agent to choose the human-like action at the current time step, but also encourages the agent to increase its chances to match the human-like actions at the future time steps.

In this project, we plan to use the same *offline* framework. Instead of using only the optimal human demonstrations to train the agent, sub-optimal demonstrations with possible factual errors will also be included. Based on the rich factuality annotations, we can explicitly design a reward function that discourage the agent to generate hallucinated text. Our whole training framework can be thought of as an *offline* training framework with sub-optimal demonstrations.

### 2.2 *Online* Training Framework

*Online* training usually use evaluation metrics directly as rewards. In abstractive summarization, reference summaries, or human demonstrations, may not be the only valid summaries of documents. To capture the translation-invariant of generated text, Paulus et al. (2017) designed a reward loss function where the objective is to directly maximize the ROUGE score (Lin, 2004, one evaluation metric of text generation) of the generated text against the reference summary. They mixed the reward loss function with the standard MLE loss function to

[2] Code available at https://github.com/yzpang/gold-off-policy-text-gen-iclr21.

encourage the fluency of the generated text. Similarly, BLEU metric (Papineni et al., 2002) is used for RL approaches to improve the quality of neural machine translation (Wu et al., 2018).

### 2.3 Learning from Human Feedback

Learning from human feedback has been explored in RL literature (Abbeel and Ng, 2004; Knox and Stone, 2010) and has been recently adopted in NLP (Ziegler et al., 2019; Böhm et al., 2019; Stiennon et al., 2020). Böhm et al. (2019) introduced a new form of reward function learned from extra data of pair-wise human evaluation score. More specifically, for each document in the dataset, different text sequences are generated based on four pre-trained summarization systems. Human evaluation scores are then provided for each pair of the text sequences to rank the generation quality. Using these ranked text sequences, a reward function for ranking can be learned supervisely. Based on this learned reward function, the RL-based summarization systems are proved to generate more human-appealing summarizations. Similar results are shown in Stiennon et al. (2020).

In comparison, we propose to (1) leverage available datasets that have fine-grained token level annotation on machine generated summaries instead of a preference score over summaries; (2) assign the reward based on the annotated summaries, and does not explicitly learn a parameterized reward function which may not generalize to the new histories generated by the RL system.

### 2.4 Factual Hallucinations

Abstractive summarization models are subject to hallucinated content that is commonly inconsistent with respect to the source. Hallucinations consists of **intrinsic hallucinations** and **extrinsic hallucinations** (Maynez et al., 2020). Intrinsic hallucination happens when models lack an understanding of the source and therefore misuse part of its information. Extrinsic hallucination is caused by adding external information that cannot be inferred from the source but may be factually correct due to the knowledge learned during model's pre-training.

**Evaluation of Factual Consistency** Although metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) are measurements of content informativeness of the generated text to some extent, there are not well-correlated with factual consistency (Falke et al., 2019; Kryscinski et al., 2020) which is a crucial aspect especially for text summarization. Models with training objectives involving these n-gram based metrics may improve the performance in terms of n-gram overlap but may fail to address the hallucination problem.

To overcome the hallucination problem in abstractive summarization, Cao et al. (2021) trained a summarization model with factuality-based rewards. Specifically, they used a trained factuality classifier to identify non-factual entities from the training set human demonstrations and assigned a negative reward to the next token (action) if it is classified as non-factual. A positive reward is assigned for the next token (action) using probability under MLE otherwise. An earlier work (Zhang et al., 2020b) combines both ROUGE reward and factuality reward in generating more fluent and faithful summaries of clinical notes. Compare to these methods, we train our model by leveraging a set of sub-optimal machine generated summaries which have rich human factuality annotations.

## 3 Task Formulation

The present work aims to improve the factual consistency of generated summary via RL in text summarization. Given pairs of source articles and human references $(D, Y)$, text summarization learns to generate a concise and informative summary $y'$ by distilling the most salient information from a new source document $d$. Mathematically, text summarization models the conditional likelihood $p(y'|d) = \sum_t p(y'_t|y'_{<t}, d)$, where $y'_{<t}$ denote generated tokens before time step $t$.

Rather than the standard training procedure that uses $(d, y)$ pairs from $(D, Y)$, we leverage a pre-trained summarization model and keep fine-tuning on it using pairs of document $d$ and SOTA-model generated summaries $y'$. In addition, we provide token-level factual consistency labels $l$ for each token in $y'$ to encourage model to not generate factually inconsistent words. In section 3.1, we describe how we construct our train and test data.

### 3.1 Dataset Construction

CNN/DM (Nallapati et al., 2016) and XSum (Narayan et al., 2018) are two typical text summarization datasets. We choose XSum for the present work as it consists of more abstractive summaries and contains more factual hallucinations compared to CNN/DM. Previous works constructed annotated summarization datasets by sampling and annotat-

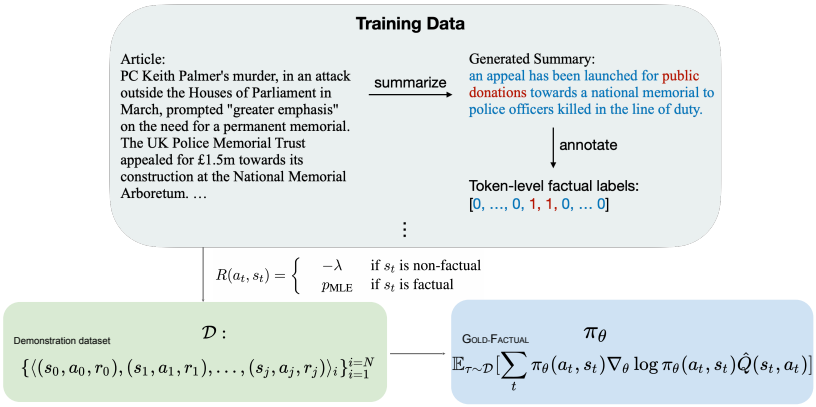


Figure 2: Framework of GOLD-FACTUAL. Text highlighted in red represents factually inconsistent span.

ing model generated summaries from the test set of XSum. We now describe several annotated datasets that we leverage in our experiments.

**CLIFF** CLIFF (Cao and Wang, 2021) has word-level factuality annotations on summaries. Each word in a summary was annotated as correct, extrinsic, intrinsic or world knowledge, where the consensus was reached by two experienced annotators. The dataset consists of 150 summaries each for both CNN/DM and XSum from BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020a) ($150 \times 2 \times 2$ summaries). Both models are state-of-the-art summarization models. We only consider the summaries generated by BART as part of our training data.

**GoyalDurrett2021** Goyal and Durrett (2021) manually identified all hallucinated text spans for each summary and classified hallucination types into {intrinsic, extrinsic} × {entity, event, noun phrase, others}. The dataset consists of 50 and 100 summaries from BART for CNN/DM and XSum, respectively.

**Cao2021** Cao et al. (2021) extracted all entities from 700 XSum summaries generated by BART and manually labeled each entities with one of the labels from {Non-hallucinated, Non-factual Hallucination, Intrinsic Hallucination, Factual Hallucination}.

**Dataset construction** To buildup our training data, we use the XSum portion of CLIFF and GoyalDurrett2021 and the entire annotated dataset from Cao2021. There are 950 examples in total in our constructed training set. Further, based on the fine-grained annotations from these datasets, we construct word-level factuality labels $l$ for each model generated summary $y'$. Specifically, we label words that are not included in any hallucinated span as factually consistent, and factually inconsistent otherwise. If any word has a hallucination label, then the entire summary is factually inconsistent, and consistent otherwise. We randomly sampled 1,000 instances $(d, y)$ from the XSum test set that do not overlap with our training dataset as our test set [3].

## 4 Method

In this section, we discuss each component of our framework (Figure 2).

### 4.1 Markov Decision process (MDP)

Text summarization can be considered as a MDP without reward or $\mathcal{M} := (S, A, T, R, \gamma)$. At each time-step $t$, the state $s_t = (y_{<t}, d) \in S$ consists of the source document $d$ and the previously generated tokens $y_{<t}$. The agent, which is the text generation model, takes an action by generating a new token $a_t$ from vocab $\mathcal{V} = A$. Depending on the action taken, the agent deterministically transitions to the next state $s_{t+1} = (y_{<t+1}, d)$ and receives a reward of $R(s_t, a_t)$. Here, $y_{<t+1}$ is the previous generated token $y_{<t}$ attached with the new token $a_t$. $\gamma$ is the discount factor. the overall objective is to find an optimal policy $\pi^* = \arg\max_\pi \mathbb{E}_{\tau\sim\pi}[J(\tau)]$, where $J(\tau)$ is the return of a trajectory $\tau$.

In the present work, we formulate the learning of text generation problem as an *offline* RL problem that assumes that the agent has the access to the trajectories generated by behavior policies, for example, human experts or other text generation models. Let

[3] Since using the entire XSum test set requires much longer time for our experiments and evaluations, we decide to only sample a subset of the test set.

$\mathcal{D} = \{\langle(s_0, a_0, r_0), (s_1, a_1, r_1), ..., (s_j, a_j, r_j)\rangle_1$ $, ..., \langle(s_0, a_0, r_0), (s_1, a_1, r_1), ..., (s_j, a_j, r_j)\rangle_{i=1}^{N}\}$, $s_j \in S, a_j \in A$, and $i, j, N \in \mathbb{N}$ be a set of demonstrated trajectories.

### 4.2 GOLD-FACTUAL

We introduce GOLD-FACTUAL an *offline* training framework similar to (Pang and He, 2021) that updates the policy by an *off-policy* policy gradient as follows:

$$\nabla_\theta J_\theta = \mathbb{E}_{\tau\sim\pi_b}[\sum_t \omega_t \nabla_\theta \log \pi_\theta(a_t, s_t)\hat{Q}(s_t, a_t)]$$

where $\omega_t = \Pi_{t'=0}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\pi_b(a_{t'}|s_{t'})}$ is the importance weight and $\pi_b$ is the behavior policy. After applying the per-action approximation that considers the sampling weight of current time step only, the importance weight becomes $\omega_t = \frac{\pi_\theta(a_t|s_t)}{\pi_b(a_t|s_t)}$. In addition, since the behavior policy $\pi_b$ is not available, $\pi_b$ is approximated by assuming all the trajectories to have equal probabilities to be sampled from the demonstration $\mathcal{D}$, and all the state-action pair are unique in its trajectories. In this way, $\pi_b$ can be approximated as $\pi_b \approx 1/N$ with $N$ the total number of trajectories in $\mathcal{D}$. After the approximations, the policy is updated by a policy gradient as follows:

$$\nabla_\theta J_\theta \approx$$
$$\mathbb{E}_{\tau\sim\mathcal{D}}[\sum_t \pi_\theta(a_t, s_t)\nabla_\theta \log \pi_\theta(a_t, s_t)\hat{Q}(s_t, a_t)]$$

### 4.3 Reward Function

Our reward function is designed to reduce the number of factual errors in a summary without sacrificing the quality of the summary. To discourage the factual error, we assign a negative non-factual penalty $-\lambda$ for each non-factual token generated in a summary. If a generated token is factual, we assign a positive reward $p_{\text{MLE}}(y'_t, y'_{<t}, d)$ that approximates human probability of generating a token $y'_t$ given the history $y'_{<t}$ and source document $d$ (the same reward function used by Pang and He (2021)). The reward function is formally defined as follow:

$$R(a_t, s_t) = \begin{cases} -\lambda & \text{if } s_t \text{ is non-factual} \\ p_{\text{MLE}} & \text{if } s_t \text{ is factual} \end{cases}$$

Considering an *offline* training scenario with possibly sub-optimal demonstration trajectories, such a reward function discourages the model to generate the same non-factual token in the demonstration, and encourages the model to generate the factual tokens as similar to human as possible. [4]

### 4.4 Evaluation Metrics

**ROUGE** We first use the standard ROUGE scores (Lin, 2004) and report the F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L, which compare the word-level unigram, bigram, and longest common sequence overlap between the generated and the human reference summary, respectively.

**Factual Consistency** We use SUMMAC-CONV (Laban et al., 2022) for factual consistency evaluation. SUMMAC-CONV is one of the state-of-the-art factuality models. Given a document and a generated summary, it outputs a score based on the aggregation of sentence-level entailment scores for each pair of input document and summary sentences. Since our goal is to evaluate whether generated summaries are factually consistent, we need to convert the scores from SUMMAC-CONV to binary labels, which requires setting up a threshold. We choose the threshold based on our training data. Specifically, we first run SUMMAC-CONV on our training set and obtain a score for each summary. We then calculate the accuracy of predicted labels using the annotated binary factuality labels under different thresholds. We choose the one that leads to the highest accuracy and directly use it to convert scores to binary labels in the test set.

## 5 Experiments and Results

In the experiments, the pre-trained Facebook models are fine-tuned for 6 epochs on the proposed training set from Section 3.1 by the GOLD-FACTUAL with non-factual penalty $\lambda = 0.1, 0.5, 1, 2, 5$. The models are then evaluated on the test set, and the ROUGE scores and factual consistency accuracy are reported in Table 1. In the remaining part of this section, we analyze the factual consistency of the text summarizations of both human and models, and investigate the trade-off between factual consistency and informativeness by sweeping the non-factual penalty $\lambda$.

[4] In the actual implementation, one word that is not in the vocab $\mathcal{V}$ can be represented by more than one tokens and we tried to only assign the non-factual penalty for the first token in a word. However, our experiments shows that this does not work as well as simply assigning the non-factual penalty for all tokens of a word. Unless stated otherwise, we assume the latter scenario.

| | Original | $\lambda = 0.1$ | $\lambda = 0.5$ | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 5$ | Human |
|---|---|---|---|---|---|---|---|
| ROUGE-1 | **42.4** | 37.4 | 37.1 | 35.8 | 33.8 | 26.9 | - |
| ROUGE-2 | **17.6** | 13.9 | 13.6 | 12.4 | 10.8 | 7.7 | - |
| ROUGE-L | **35.3** | 30.4 | 30.3 | 29.0 | 27.3 | 21.8 | - |
| FC accuracy % | 49.3 | 43.8 | 47.0 | 47.1 | 49.7 | **60.2** | 54.2 |

Table 1: The three ROUGE scores and the factual consistency accuracy of the summaries by generation models trained by GOLD-FACTUAL with non-factual penalty $\lambda = 0.1, 0.5, 1, 2, 5$ and, by the original Facebook model.

### 5.1 Factual consistency of human reference

As shown in Table 1, human reference summaries contains numerous factual hallucinations. This observation agrees with the findings from Maynez et al. (2020), where they manually annotated XSum summaries and found that only a small portion of human reference summaries have no hallucinations. This is due to the construction of the XSum dataset, where the first sentence of each news article is chosen as the summary of the article and those "summaries" commonly contains external information that is not explicitly mentioned in the article.

### 5.2 Effects of hyperparameters

As we raising the non-factual penalty factor $\lambda$ for each non-factual token generated in a summary, the F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L monotonically decreases in general and the quality of those summaries in terms of factual consistency improves and even outperforms reference summaries. Since our model receives the MLE reward $p_{\text{MLE}}(y_t, y_{<t}, d)$ by using the model generated text instead of human reference, we encourage the model to assign higher probabilities to model generated summaries. This would lead to the decrease of the ROUGE scores on the test set. As the penalty $\lambda$ increases, our model fails to maintain its capability of generating human-like summaries while out-weighting the importance of generating factual text.

### 5.3 Human Evaluation

To understand how the reward function affects the quality of generated summaries in terms of ROUGE and factual consistency, we conducted human evaluation as well. We sampled 20 examples from the test set and compared the generated summaries from the original model and our model trained with penalty $\lambda = 5$. The result is shown in Table 2. In around half of examples (55%), our model generates alternative and acceptable summaries compared to the summaries from original model. For example, in the third row of Table 2, all summaries discussed the earthquake in Japan but with unique but verifiable details about the earthquake. Notice that these alternative but valid summaries may not overlap as many with the human reference summary, which leads to lower ROUGE scores. Our model is also capable of generating factual consistent summaries that were previously factually inconsistent. For example, in the first row of Table 2, the original model generates *"the number of people arrested for drink-driving at UK airports has more than doubled in the past year"*. The actual sentence discussing this phenomenon in the source article is *"A total of 387 people were arrested between February 2016 and February 2017 - up from 255 the previous year"*. Apparently the number of arrested peopled are not doubled. Our model instead generates that the number of arrested people is raising. Although this is not as detailed as what is claimed in the human reference summary *have risen by 50%*, it is a factual consistent summary.

However, in several cases (30%) our model generates worse summaries as well. A representative example is in the second row of Table 2, where our model generates a ungrammatical sentence with lots of repetitions. The repetition problem is more common in the earlier language models but less frequent recently due to the raise of powerful large transformer-based language models. The high proportion of such repetitive summaries generated by our model may caused by our designed reward function, where our model may find a trick of obtaining higher factual rewards but down-weight its capability of generating grammatical sentences.

## 6 Limitations

There are a few limitations of our work. First, we only have limited amount of token-level annotated BART summaries for training. However, our ex-

| | Reference | Original | $\lambda = 5$ |
|---|---|---|---|
| Right Correction (10%) | Arrests of passengers suspected of being drunk at UK airports and on flights have risen by 50% in a year, a Panorama investigation has revealed. | The number of people arrested for drink-driving at UK airports has more than doubled in the past year, according to figures obtained by BBC Panorama. | The number of people arrested at UK airports on suspicion of drinking alcohol in duty-free shops has risen, according to police figures obtained by Panorama. |
| Repetition (30%) | BHP Billiton and Vale have agreed a deadline of 30 June to consolidate and settle claims resulting from Brazil's Samarco dam disaster in 2015. | The owners of Brazil's Vale and BHP Billiton have agreed a deal with prosecutors over the collapse of a dam at their Minas Gerais mine in 2014. | The companies that run BHP and other script script script scripts script script Script script script screenplay script scriptscript script ... |
| Alternative Summaries (55%) | A more powerful earthquake has rocked the southern Japanese city of Kumamoto in the middle of the night, a day after an earlier tremor killed nine people. | A powerful earthquake has struck south-east Japan, hours after a powerful quake struck the same area on Thursday, causing extensive damage and chaos. | A huge earthquake has hit a region of Japan, similar to that hit the country on Thursday night, which sparked a huge tsunami and nuclear meltdown. |
| Not Main Idea (5%) | The government has rejected an online petition, signed by more than 4.1 million people, calling for a second EU referendum to be held. | More than 100,000 people who signed a petition calling for a second referendum on the UK's EU membership have been told by the Foreign Office it will not be considered for debate by MPs. | A petition calling for a re-run of the EU referendum if the turnout for the vote for Brexit was "less-than-expected" has been submitted to the Foreign Office. |

Table 2: Comparison of four types of representative examples between human reference, summaries generated by the original model, and by our model trained with penalty $\lambda = 5$. Factuall inconsistent spans are highlighted in red. The source article is not shown in the table due to the length constraints.

periments shows the promising directions of using sub-optimal summaries to improve the factual consistency. Second, the current evaluation of factual consistency relies on an imperfect factuality system and a small scale error analysis. A larger-scale of human evaluation, such as the number of mislabeled examples, could bring more insights. Third, we finally raises the degree of factual consistency of generated summaries, but this also leads to some problems such as ungrammatical text and less informative summaries.

## 7 Conclusion

In this study, we propose a new training framework GOLD-FACTUAL for text summarization to addresses the hallucination problem in abstractive summarization. We systematically analyze the performance of GOLD-FACTUAL in terms of ROUGE and a factual consistency metric. We show that our framework has the potential to generate better summaries compared to human reference summaries. A small-scale human evaluation shows that there is still a trade-off between factual consistency and other factors that measures the quality of summaries. Lastly we point out several limitations of our work and shed light on the future directions in leveraging sub-optimal summaries to improve the factual consistency.

## References


Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA. Association for Computing Machinery.

Florian Böhm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2021. Inspecting the factuality of hallucinated entities in abstractive summarization. *arXiv preprint arXiv:2109.09784*.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of NAACL*.

Ferenc Huszár. 2015. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*.

W.B̃radley Knox and Peter Stone. 2010. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*.

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

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. scpSummaC/scp: Revisiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

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

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.

Richard Yuanzhe Pang and He He. 2021. Text generation by learning from demonstrations. In *International Conference on Learning Representations*.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

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

Yuhao Zhang, Derek Merck, Emily 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*, pages 5108–5120, Online. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.