Gradient-based Adversarial Factual Consistency Evaluation for Abstractive Summarization

Zhiyuan Zeng[1]*, **Jiaze Chen**[2], **Weiran Xu**[1], **Lei Li**[3]*
[1]Beijing University of Posts and Telecommunications, Beijing, China
{zengzhiyuan,xuweiran}@bupt.edu.cn
[2]ByteDance AI Lab [3]University of California, Santa Barbara
chenjiaze@bytedance.com lilei@cs.ucsb.edu

Abstract

Neural abstractive summarization systems have gained significant progress in recent years. However, abstractive summarization often produce inconsisitent statements or false facts. How to automatically generate highly abstract yet factually correct summaries? In this paper, we proposed an efficient weak-supervised adversarial data augmentation approach to form the factual consistency dataset. Based on the artificial dataset, we train an evaluation model that can not only make accurate and robust factual consistency discrimination but is also capable of making interpretable factual errors tracing by backpropagated gradient distribution on token embeddings. Experiments and analysis conduct on public annotated summarization and factual consistency datasets demonstrate our approach effective and reasonable. Our codes can be found at https://github.com/parZival27/GrAdualCC

1 Introduction

Text summarization aims to produce a simplified version of the source document while retaining salient information. Abstractive summarization is a branch of methods in which generation text is free from constraint on the tokens that appeared in the source. These methods are extensively studied since its flexibility and generalization ability(See et al., 2017; Paulus et al., 2018; Gehrmann et al., 2018; Dong et al., 2019a). However, a challenge in abstractive summarization is the trade-off between abstractiveness and factual consistency. Recent studies show that about 30% of the summaries generated by abstractive models contain facts errors toward source documents. The proportion will rise further as the data abstractiveness increases(Cao et al., 2018; Durmus et al., 2020; Kryscinski et al., 2020), causing factual checking an essential process to verify the credibility and usability of models.

article:
The Swift Archway Cranford 545 caravan was stolen from a site in Yaxley, Cambridgeshire, on Thursday night. Davis tweeted "My touring caravan was stolen.. even though it was locked up with hitch & wheel lock!". (...) Davis has played the role of Professor Flitwick in the Harry Potter films and Nikabrik in The Chronicles of Narnia: Prince Caspian. (...)
claim:
A caravan locked by Harry has been stolen from a site in cambridgeshire.
reference:
A caravan locked by Davis has been stolen from a site in cambridgeshire.

Table 1: An incorrect summary generated by XSUM.

In Table 1, we propose an inconsistent generation example, where the blue part support factual consistency and the red part leads to factual errors.

Previous approaches for detecting or boosting factual consistency can be divided into three kinds. (1) Employ information extraction tools to extract facts and leveraging it by building additional objective (Cao et al., 2018; Goodrich et al., 2019; Zhang et al., 2020a; Zhu et al., 2021). (2) Use natural language inference or question answering models for fact checking correction (Li et al., 2018; Falke et al., 2019; Wang et al., 2020; Dong et al., 2020; Durmus et al., 2020; Chen et al., 2020). (3) Train a factual consistency evaluation model on artificial datasets generated by rule-based transformations(Kryscinski et al., 2020; Cao et al., 2020). Most of the above approaches focus on factual consistency *evaluation*. Some of these explore using pretrained language models (Devlin et al., 2019; Dong et al., 2019b; Lewis et al., 2020) to make an end-to-end fact *correction*. However, fact correction through text generation may further increase uncertainty. By comparison, the *tracing* of factual errors by explicitly marking out the latent inconsistent tokens in the generated summaries can provide more reliable and interpretable information. It has significant meaning in a real scenario but attracts less research attention.

*Work is done while at ByteDance.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4102–4108
November 7–11, 2021.

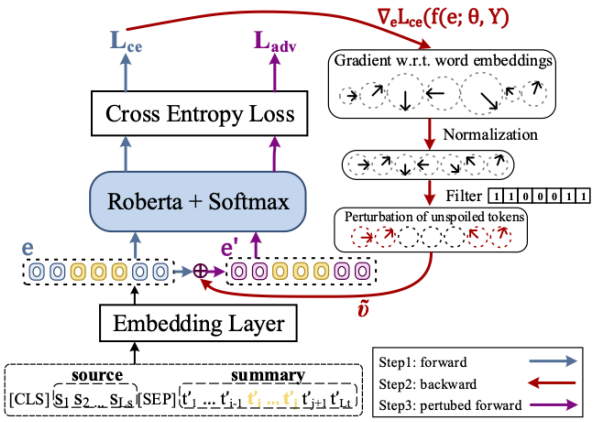

Figure 1: The overall architecture of our proposed method and a training process for inconsistent example.

In this paper, we propose a robust weakly-supervised factual consistency evaluation model and gradient-based factual errors tracing strategy. Specifically, we construct artificial datasets based on benchmark summarization datasets to train the model. Except rule-based transformations proposed by (Kryscinski et al., 2020; Cao et al., 2020), we propose an implicit augmentation to obtain hard factual inconsistent examples by the adversarial attack. It alleviates the problem of oversimplified negative samples and therefore improves the model performance and robustness. Further, we propose a novel strategy to trace factual errors based on gradients distribution without adding any parameters. The analysis on gradients also provides stronger interpretability for the factual consistency evaluation results. Our contributions are three-fold: (1) We propose an efficient adversarial data augmentation approach to generate weakly supervised samples for factual consistency evaluation. (2) We design a novel strategy to tracing factual errors by utilizing gradient distribution. (3) Experiments and analysis conducted on various datasets show the effectiveness and interpretability of our proposed methods.

2 Methods

Fig 1 shows the overall architecture of our factual consistency evaluation model. We adopt Roberta(Liu et al., 2019) as feature extractor $f(\cdot)$. Given a sequence concatenated by source document and corresponding summary $\boldsymbol{x} = \{x_1, x_2, \ldots, x_{L_x}\}$, we encode tokens into representations $r_i = f(E(x_i))$, where $e_i = E(x_i)$ indicates the embedding process. We add a simple linear layer after representation of [CLS] token for calculating binary cross-entropy loss $\mathcal{L}_{ce}(f(\boldsymbol{e}; \theta, Y))$, where $Y \in \{Consistent, Inconsistent\}$.

In the following, we describe the building methods of artificial datasets, the model's training process, and our proposed error tracing strategy.

Artificial Dataset. We follow the recent methods to generate inconsistent samples through rule-based transformations (Kryscinski et al., 2020; Cao et al., 2020). The source document $\boldsymbol{s}$ and the corresponding target summary $\boldsymbol{t}$ is treat as a consistent example $\boldsymbol{x}_p = \{\boldsymbol{s}, \boldsymbol{t}\}$. After utilizing corruption on part of the original summary (corresponding to yellow highlighted part in Fig. 1), the source and the pseudo summary $\boldsymbol{t}'$ forms a inconsistent example $\boldsymbol{x}_n = \{\boldsymbol{s}, \boldsymbol{t}'\}$.

Three types of strategies are used to corrupt the original summaries. (1) *Entity swapping*[1]: replace a random entity in the reference summary by another random entity of the same type in the same source document. To alleviate the bias caused by synonyms, we apply an empirical threshold on the similarity between the original and pseudo entities based on the simple distance algorithm. (2) *Pronoun swapping*, replace a random pronoun with another one of matching syntactic case. (3) *Negation*: transform a random auxiliary verb to its negative form.

Adversarial Augmentation. It is pointed that a classifier trained on artificial datasets only works well on easy examples, thus can hardly generalize well to actual scenarios (Zhang et al., 2020b). To alleviate this, we propose an adversarial attack mechanism (Goodfellow et al., 2015; Kurakin et al., 2016; Miyato et al., 2017; Yan et al., 2020; Meng et al., 2020) on rule-based pseudo samples as data augmentation. For token embeddings of a sample $\boldsymbol{e}$, we try to find a worst-case perturbation vector $\widetilde{\boldsymbol{v}}$ that maximizes the loss function:

$$\widetilde{\boldsymbol{v}} = \underset{||\boldsymbol{v}|| \leq \epsilon}{\arg\max}\, \mathcal{L}_{\text{ce}}(f(\boldsymbol{e} + \boldsymbol{v}; \theta), Y) \quad (1)$$

Where ϵ is the norm bound of the perturbation, since the complexity of neural models, it is intractable to compute the perturbation precisely. Instead, we apply Fast Gradient Value (FGV) (Rozsa et al., 2016) to approximate a worst-case perturbation:

$$\widetilde{\boldsymbol{v}} = \epsilon \frac{g}{||g||}; \text{where } g = \nabla_{\boldsymbol{e}} \mathcal{L}_{\text{ce}}(f(\boldsymbol{e}; \theta), Y) \quad (2)$$

The gradient g is the first-order differential of

[1]We extract entities by pre-trained NER model of spaCy https://spacy.io/.

Model	CNN/DM				XSUM			
	Reference	Factual Annotation			Reference	Factual Annotation		
	acc	acc	bacc	f1	acc	acc	bacc	f1
FactCC	53.5	86.1	72.7	70.6	59.7	73.0	54.1	51.8
FactCCX	54.3	**86.5**	72.9	**71.1**	57.1	60.0	53.1	50.3
FEC	79.6	83.5	66.5	66.5	90.2	74.1	56.7	52.7
ours w/o adv.	81.1	85.6	72.5	68.4	90.1	73.9	56.7	52.5
ours	**83.3**	86.0	**73.3**	70.1	**90.7**	**74.7**	**57.2**	**53.4**

Table 2: Performance comparison between our method and baselines on CNN/DM and XSUM datasets.($p<0.05$).

$\mathcal{L}_{\text{ce}}$, representing the direction that rapidly increases the loss. We normalize the gradient and use a small norm to ensure the approximation reasonable. Specifically, for inconsistent samples, the perturbations on the corrupted tokens are masked by a filter. We will explain the reason in the following. It is worth noting that Fig. 1 only shows an inconsistent example, and for consistent examples, perturbations on all tokens are retained. We name the filtered perturbation as $\widetilde{\boldsymbol{v}}_{\text{p}}$ and add it to $\boldsymbol{e}$ to obtain new tokens embeddings $\boldsymbol{e}' = \boldsymbol{e} + \widetilde{\boldsymbol{v}}_{\text{p}}$, which can be regarded as an augmented sample. We feed it to the model again to obtain another loss $\mathcal{L}_{\text{adv}}$ with the same label. Finally, we use the weighted sum of two losses to train our model:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{ce}} + (1 - \alpha) \cdot \mathcal{L}_{\text{adv}}$$

Error Tracing. We propose a novel factual error tracing strategy using back-propagated gradients g. Instead of introducing more neural network layers and parameters, our proposed method can be regarded as an inherent by-product of adversarial augmentation.

Let $\Delta\mathcal{L} = \mathcal{L}_{adv} - \mathcal{L}_{ce} \geq 0$, the overall loss can be simplify as: $\mathcal{L} = \mathcal{L}_{ce} + (1 - \alpha) \cdot \Delta\mathcal{L}$. With the optimization of the model, $\Delta\mathcal{L}$ tends to zero, which reduces the loss change caused by adversarial perturbations, so that the representations of perturbed tokens tend to remain relatively stable in its neighborhood of the high-dimensional space. For inconsistent samples, as perturbations of corrupted tokens are masked, these tokens retain sensitiveness to loss change. So gradient will show a relatively higher value in the corrupted tokens as the loss changes faster when this part changes.

Generalizing the phenomenon into the test process, the model can use gradients distribution to trace factual errors. The cross-entropy loss is back-propagated to the embedding layer to obtain a gradient distribution. We use top-k algorithms to filter candidate error tokens on samples with inconsistent predictions. We conduct quantitative analysis and visualization in Section 4 to demonstrate the effects.

3 Experiments

3.1 Experimental Setup

We perform experiments on two benchmark text summarization datasets CNN/DM(Nallapati et al., 2016) and XSUM(Narayan et al., 2018). Weakly supervised training data was generated as described in Section 2. Models are evaluated in two ways: (1) with the source documents and the ground truth references of datasets, which are all positive examples (2) with the manual factual consistency annotations provided on CNN/DM(Kryscinski et al., 2020) and XSUM(Maynez et al., 2020). We report accuracy, balanced accuracy, and marco F1-score.

We compare our evaluation model with strong baselines: (1) FactCC(Kryscinski et al., 2020): a BERT-based classification model trained on artificial datasets. (2)FactCCX(Kryscinski et al., 2020): a version of FactCC with additional span selection heads. (3) FEC(Cao et al., 2020): a BART-based factual error evaluation and correction model trained on artificial datasets.

3.2 Datasets details

The artificial dataset constructed on CNN/DM contains 408369 examples, in which 200000 examples are factual consistent, and 208369 examples are inconsistent. The artificial dataset constructed on XSUM contains 608262 examples, in which 300000 examples are factual consistent, and 308262 examples are inconsistent. CNN/DM factual consistency annotation dataset (Kryscinski et al., 2020) contain 441 consistent samples and 62 inconsistent samples. XSUM factual consistency annotation dataset (Maynez et al., 2020) contain 199 consistent samples and 1667 inconsistent samples. We split the artificial datasets into the training set, the development set, and the test set in a ratio of 90%, 5%, and 5%.

article: (...) Racing, watched by new Indian owner Ahsan Ali Syed, took the lead in the 33rd minute through Argentine striker Ariel Nahuelpan. Valencia were reduced to 10 men when defender David Navarro was booked in the 54th minute. (...) ***claim***: (...) Defender Ariel Nahuelpan is sent off in second half before Tino Costa equalizes. Both teams then have a player sent off at the end of the match. (...)
article: (...) When she arrived in Vancouver, however, she says she was required to work 16 hours a day, seven days a week, with no days off and no statutory holidays. (...) Things came to a head in June 2010 when Sarmiento called the police after getting into a confrontation with Huen. (...) ***claim***: (...) Court was told how the maid worked 16 hours a day for June 2010 without a holiday. (...)
article: (...) "These degenerate molesters are cowards,"Timothy J. McGinty said. "... This man couldn't take, for even a month, a small portion of what he had dished out for more than a decade." (...) ***claim***: Prosecutor: Castro could take what he dished out for a decade. (...)

Table 3: Article fragments and corresponding claims on the artificial test set constructed on CNN/DM.

3.3 Implementation details

We finetune our model on public pre-trained model Roberta-base (Liu et al., 2019), which has 12 layers, 768 hidden states, and 12 heads. The max length of the input is 512. Adam is used for optimization with an initial learning rate of 1e-5, and the batch size is 16. We set the training epoch up to 3 with evaluation on the validation set every 1000 steps. The range of weight between two losses is 0 to 1. We empirically set $\alpha = 0.5$ to adapt to the general situation. The amplitude of adversarial perturbation is obtained by the heuristic method in the range of 2E-3 to 1E-2. Within the range, the influence of amplitude on model performance is less than 3%, and 6E-3 gain the best performance. Each result of the experiments is tested five times under the same setting and gets the average value. The training stage of our model lasts about 2.0 hours per epoch on four pieces of Tesla-V100-SXM2(32GB). The average value of the trainable model parameters is 476M.

3.4 Main Results

Table 2 shows the main results. Our evaluation model gains higher accuracy on both datasets' ground truth references, which are significantly better than FactCC and FactCCX. Since our model corrupts the reference summary rather than a fragment of source document, it fits better with abstractive summarization. On factual consistent dataset of CNN/DM, our model significantly outperform FEC by 2.5%(accuracy), 6.8%(balance accuracy), and 3.6%(marco-F1), and shows close results with the previous state of the art model FactCCX. On XSUM, our model gains consistent improvement on all metrics. However, every model performs poorer on XSUM than CNN/DM, indicating that higher abstractiveness makes fact consistency evaluation more difficult. Besides, we conduct an ablation experiment on adversarial augmentation. The result shows that implicitly augment data through adversarial attack significantly benefits the evaluation, and the improvement on CNN/DM is more pronounced. It confirms that the rule-based augmented data can only simulate simple situations. In general, our proposed evaluation model is more reasonable for the factual consistency evaluation of abstractive text summarization.

	k	1	2	3	4	5
token	w/o. adv.	50.3	68.2	78.5	83.7	86.8
	w. adv.	54.1	70.7	80.1	84.8	87.5
	Relative↑	7.55%	3.67%	2.04%	1.31%	0.81%
span	w/o. adv.	33.0	53.8	68.2	76.0	80.5
	w. adv.	36.5	56.1	70.1	77.2	81.4
	Relative↑	10.61%	4.28%	2.76%	1.58%	1.12%

Table 4: Recall of errors under different settings.

4 Analysis

4.1 Case Study.

Table 3 shows cases study of error tracing. We display some inconsistent samples of the artificial test set construct on CNN/DM. For the original text, the blue highlighted part represents the original span appearing in the ground truth (if it has). The orange highlighted part represents the pseudo span used to corrupt the ground truth. We normalize the predicted gradient distribution and use varying degrees of red to describe the tokens with top-5 gradient values. The brighter red represents the larger gradient.

We found that gradients show a high value on the corrupted part. It indicates that our method can provide instructive error tracing results and robust to different error types. Further, The analysis of gradient distribution explicitly explains the factual consistency evaluation result, which improves the interpretability of prediction results.

4.2 Quantitative Analysis.

Table 4 shows the quantitative results of our error tracing methods. We collect gradient distribution

Source article fragments		
(...) Creams such as Arnicare for pain relief or liquids suchas Sidda Flower Essences for male virility are part of a $2.9 billion business that has seen "explosive growth," according to the FDA. These drugs do not go through the same level of scrutiny as over-the-counter and prescription drugs. (...)	(...) Rabbis Mendel Epstein, 69; Jay Goldstein, 60; and Binyamin Stimler, 39, were found guilty on one count of conspiracy to commit kidnapping in New Jersey federal court. Goldstein and Stimler were also convicted on charges of attempted kidnapping. (...)	(...) In his remarks at an anti-vaccination movie screening, he decided to compare "vaccine-induced" autism to the Holocaust. He said, "They get the shot, that night they have a fever of a hundred and three, they go to sleep, and three months later their brain is gone," Kennedy said. (...)
Model generated claims		
Drugs do not go through the same level of scrutiny as over-the-counter and prescription drugs.	Rabbis mendel epstein, 69, and binyamin stimler, 39, were found guilty on one count of conspiracy to commit kidnapping in new jersey federal court.	He decided to compare "vaccine-induced" autism to the holocaust, he says.

Table 5: Inconsistent cases on CNN/DM factual consistency annotation dataset.

on token embeddings of inconsistent samples in CNN/DM artificial test set and treat the tokens with top-k gradient values as predictive factual errors. For a range of k, we compute token recall and span recall, where the token recall allows only predicting the portion of the errors and the span recall requires including all error tokens. We treat the model without adversarial augmentation as a baseline.

The results indicate that with adversarial augmentation, the performance of error tracing gains consistent improvement on both token level and span level. When $k = 3$, more than 70% of spans and more than 80% of tokens can be recalled. When k is smaller, the recall improvement caused by adversarial augmentation is relatively significant. Besides, although span recall has lower metrics due to stricter restrictions, the method we propose can achieve a greater relative improvement. In summary, we have proved that (1) effective error detection can be carried out through gradient distribution (2) our proposed adversarial augmentation can optimize gradient distribution.

4.3 Error analysis

Table 5 shows some inconsistent cases on CNN/DM factual consistency annotation dataset (Kryscinski et al., 2020), which our model make wrong prediction. The blue part represents the content covered by the claim, and the red part denotes the content claim neglect or makes the wrong expression. We found that these examples all overlap with the source text a lot, and errors occur only in very small parts. This is consistent with FactCC's data structure strategy, but its universality in abstractive summarization tasks needs further study.

Model	Error Tracing Helpfulness			Correlation with Human
	Helpful	Somewhat Helpful	Not Helpful	
ours w/o adv.	78.04%	14.60%	7.36%	0.722
ours	83.86%	10.34%	5.80%	0.758
oracle	94.10%	4.16%	1.74%	0.953

Table 6: Human evaluation on artificial dataset.

4.4 Human Evaluation.

Table 6 shows the human evaluation results of our models on CNN/DM artificial dataset. Following (Kryscinski et al., 2020), we randomly sample 500 data pairs in the artificial test set and highlight the tokens with the top-5 predicted gradient value. The staff judges the factual consistency and gives whether the highlighted content provides help for the judgment. In addition, we compute a Pearson correlation between the human factual consistency judgment results and the predicted label of the model. $oracle$ means using ground truth labels and corrupted span, which set an upper bound for evaluation. The results show that the adversarial mechanism significantly improves the availability of error tracing information and evaluation performance.

5 Conclusion

Abstractive summarization models are susceptible to factual inconsistency generation. To optimize the robustness and interpretability of factual consistency evaluation, we proposed an implicit data augmentation method based on the adversarial attack to construct hard factual inconsistent examples and gradient-based fact errors tracing strategy to provide auxiliary information. Experiments conduct on public datasets demonstrate the effectiveness of our models. The extensive analysis further reveals the role of the error tracing strategy.

6 Broader Impact

Abstractive summarization systems have demonstrated remarkable performance across a wide range of applications, with the promise of a significant positive impact on human production mode and lifeway. However, due to excessive abstractiveness, current models usually face an unfaithful generation problem, which may affect human judgment and impair the safety of models in practical applications, thus severely restricts the development of technology. In domains with the most significant potential for societal impacts, such as news, models should recognize factual errors to avoid bad influence. Our work focuses on the robustness and interpretability of the factual consistency evaluation model to take a step towards the ultimate goal of enabling the safe real-world deployment of abstractive summarization systems.

References

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. 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.

Jiangjie Chen, Qiaoben Bao, Jiaze Chen, Changzhi Sun, Hao Zhou, Yanghua Xiao, and Lei Li. 2020. Loren: Logic enhanced neural reasoning for fact verification. *ArXiv*, abs/2012.13577.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019a. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019b. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

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

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

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

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. 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*.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 166–175. ACM.

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.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

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

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

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

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.

Kevin Meng, Damian Jimenez, Fatma Arslan, Jacob Daniel Devasier, Daniel Obembe, and C. Li. 2020. Gradient-based adversarial training on transformer networks for detecting check-worthy factual claims. *ArXiv*, abs/2002.07725.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Andras Rozsa, Ethan M Rudd, and Terrance E Boult. 2016. Adversarial diversity and hard positive generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–32.

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

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

Yuanmeng Yan, Keqing He, Hong Xu, Sihong Liu, Fanyu Meng, Min Hu, and Weiran Xu. 2020. Adversarial semantic decoupling for recognizing open-vocabulary slots. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6070–6075, Online. Association for Computational Linguistics.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020a. 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.

Yuhui Zhang, Yuhao Zhang, and Christopher D. Manning. 2020b. A close examination of factual correctness evaluation in abstractive summarization.

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