Improving Factual Consistency of Abstractive Summarization on Customer Feedback

Yang Liu, Yifei Sun, Vincent Gao
Amazon, Inc.
{yngliun, sunyifei, vincegao}@amazon.com

Abstract

E-commerce stores collect customer feedback to let sellers learn about customer concerns and enhance customer order experience. Because customer feedback often contains redundant information, a concise summary of the feedback can be generated to help sellers better understand the issues causing customer dissatisfaction. Previous state-of-the-art abstractive text summarization models make two major types of factual errors when producing summaries from customer feedback, which are *wrong entity detection* (WED) and *incorrect product-defect description* (IPD). In this work, we introduce a set of methods to enhance the factual consistency of abstractive summarization on customer feedback. We augment the training data with artificially corrupted summaries, and use them as counterparts of the target summaries. We add a contrastive loss term into the training objective so that the model learns to avoid certain factual errors. Evaluation results show that a large portion of WED and IPD errors are alleviated for BART and T5. Furthermore, our approaches do not depend on the structure of the summarization model and thus are generalizable to any abstractive summarization systems.

1 Introduction

In order to improve customer order experience, most e-commerce stores allow customers to submit reviews or feedback via their post-order communication channels. Such customer feedback, usually in the form of short paragraphs of free texts, contains information reflecting the issues that customers experienced in their purchases. This information can be shared with sellers to bring their awareness on the problems in their products. However, customer feedback often include other contents that are irrelevant to the product issues. Such redundant information requires extra efforts for sellers to fully understand the customers major concerns, and sometimes even causes confusion.

Source: (...) I ordered this mouse for my new laptop. However, when I received it, I could see many scratches on the product. It looks like it has been used before. (. . .) **Reference Summary**: The mouse delivered has many scratches. It looks like it has been used. **Model Summary**: The laptop came with many scratches, looks like it has been used.
Source: (. . .) I checked the serial number and found it doesn't match the one on the website. This phone is not defective. I question the source of this product (. . .) **Reference Summary**: The phone serial number doesn't match the one on the website but the phone is not defective. **Model Summary**: This phone is defective and the serial number doesn't match the one on the website.

Table 1: Examples of the two major factual errors: WED (upper) and IPD (lower).

To reduce the redundancy, a concise summary of customer feedback can be provided where the information is concentrated on the product issues while other irrelevant contents are filtered out. Such summary allows sellers to quickly capture and comprehend the problems, and thus they can address buyer dissatisfaction more efficiently.

The problem of generating summaries from customer feedback is modeled as a text summarization task (Nallapati et al., 2016; Allahyari et al., 2017; Gao et al., 2020) in the natural language processing (NLP) domain. Abstractive summarization models with transformer-based architecture have achieved success in a variety of summarization tasks (Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020; Bao et al., 2020). Hence, we harnessed the recent state-of-the-art (SOTA) abstractive summarization models, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), and fine tuned the models for our specific summarization task. We aim to utilize summarization models to produce the summary that can correctly describe the product issues presented in customer feedback. However, from human evalu-

ation results, we observed that the summary generated by these abstractive summarization models sometimes contains the information that is inconsistent with facts in the input text. Such factual inconsistencies have also been observed in previous studies (Cao et al., 2018; Kryscinski et al., 2019, 2020). More specifically, we analyzed 75 inconsistent summaries obtained from human evaluations on more than 600 model-generated summaries. We found that around 70% factual inconsistent summaries [1] follow two error patterns: *wrong entity detection* (WED) and *incorrect product-defect description* (IPD). The error of WED often occurs in the cases where the feedback text involves multiple entities but the models fail to detect the primary entity. For IPD, the generated summary contains the product-defect description that contradicts with the original description in the customer feedback. Table 1 shows the examples [2] of the two types of factual errors.

In this work, we propose a set of methods in order to improve the factual consistency of abstractive summarization on customer feedback. We first introduce specific factual errors into each target summary to generate their negative counterpart. We then use such pair of consistent and inconsistent summaries with a contrastive loss term added in the training objective to enhance the model's robustness against the two major factual errors.

Our contributions are two folds. First, The proposed approaches with corrupted summary generation and contrastive loss augmentation do not pose requirements on the achitecture of the summarization model. Thus, they can be applied to any abstraction-based summarization model to improve the model faithfulness. Second, we test the proposed approaches on SOTA summarization algorithms such as BART and T5. Our approaches show large benefits in reducing the common factual errors in customer-feedback summarization.

2 Related Work

There have been increasing research attentions on improving the factual consistency of abstractive summarization models. Lots of priors work focused on different ways of adding external signals or constraints to enhance the summary generation. Cao et al. (2018) built a dual-attention framework so that the summary generation is conditioned on both the source document and extracted key information. Li et al. (2018) incorporated the entailment knowledge by utilizing entailment-aware encoder and decoder. With using the textual entailment, Falke et al. (2019) re-ranked the candidates summaries to select the summary that's better aligned with the source document. Dou et al. (2020) studied different external signals, including key sentences, keywords and relations, and used them in addition to the input text to guide the summary generation. Mao et al. (2020) constrained certain tokens to require them to be present in the summary. Similarly, Yuan et al. (2020) add constraints on the model to include certain attribute words in the product summarization. Zhu et al. (2021) integrated information extraction and graph attention network into transformer-based seq2seq framework.

To identify and correct the unfaithful summaries, Wang et al. (2020) proposed to use a question answering framework to check the faithfulness of the summary while Dong et al. (2020) built a factual correction model that leverages knowledge learned from question answering models. Kryscinski et al. (2020) trained a BERT-based model to classify whether the summary is factual consistent. Cao et al. (2020) and Zhu et al. (2021) developed factual corrector based on BART (Lewis et al., 2020) and UniLM (Dong et al., 2019), as a post-processor to rectify factual errors from the upstream summarization model. They corrupted the reference summaries with artificial errors and used them as the negative samples for training the correctors. In our work, we also generate corrupted summaries as the negative counterparts of the target summaries. The difference is that, instead of building a separate corrector model, we directly engineer the training objective of the summarization model. By leveraging contrastive learning (Schroff et al., 2015; Khosla et al., 2020), we define contrasive losses to guide the output summary away from certain factual errors.

3 Proposed Approaches

Our error analysis of customer-feedback summarizaton showed that most of the factual errors belong to two error types: WED and IPD. Hence, in our proposed approaches, we first apply rule-based transformations and introduce synthetic factual errors of the two error patterns into the target summaries. We then modify the training objective by

[1] The rest of the unfaithful summaries are due to miscellaneous factual errors that are hard to cluster.

[2] Due to confidentiality, all customer feedback examples in this paper are composed by the authors.

Source: (...) I've bought cheese from this store for many times, and they were very good. So I think other products must be good too. Then I ordered several bottles of milk. But they are clearly expired (. . .) **Reference Summary**: Milk delivered is expired. **Corrupted Summary**: Cheese delivered is expired.
Source: (. . .) The eggs I purchased have bad smells. They don't look like fresh eggs. (. . .) **Reference Summary**: Eggs have bad smells, and don't look like fresh eggs. **Corrupted Summary**: Eggs have good smells, and don't look like fresh eggs.

Table 2: Examples of corrupted summaries. We replace the primary entity in the first example and switch the description in the second example.

adding the contrastive loss so as to guide the model to avoid those mistakes.

3.1 Synthetic Factual Errors

We augment the training data by applying two types of corruption methods on the target summary. The corruptions are designed to mimic the factual errors we observed. In the first method, we replace the named entities in the target summary with the other random entities of the same type in the source document. If no such replacement entity can be found in the source document, we randomly pick one from the top 50 appeared entities in our dataset. We used Spacy toolkit (Honnibal et al., 2020) for the named entity extraction. In the second method, we use predefined rules to transform the product-defect description in the target summary. We detect the adjectives describing the product defect and switch their sentiment. There are two ways that we change the description. One is by adding negation word *not* before the adjective. For example, we alter *"product is broken"* to *"product is not broken"*. If word *not* is already presented, we will remove it instead. The other way is by switching a descriptive word to the one with opposite meaning, such as changing *"opened"* to *"sealed"*. Table 2 shows some examples of the corrupted summaries.

3.2 Training Objective

For each training sample, we now have a triplet (d, s_+, s_-) consisting of the source document d, target summary s_+, and corrupted summary s_-. The summarization model takes d as the input and generates the output o. Our training objective is to drive the model output o to resemble s_+ while at the same time avoiding the factual errors presented in s_-. Inspired by contrastive learning (Schroff et al., 2015; Khosla et al., 2020), we compare different contrastive loss functions for model training.

Direct Contrast Compared to the ordinary loss function for summarization, we add an extra term that takes into account the informration from corrupted summary:

$$\mathcal{L}_{DC} = \mathcal{L}(s_+, o) - \alpha * \mathcal{L}(s_-, o)$$

where $\mathcal{L}(s_+, o)$ is the cross entropy loss between s_+ and o, $\mathcal{L}(s_-, o)$ is the cross entropy loss between s_- and o, and α is a tunable hyperparameter controlling the impact from the second term. The loss function will purely focus on the difference between s_+ and s_- if $\alpha = 1.0$. Thus, we generally use small value for α to ensure the model will produce fluent summary.

Constrained Negative Here, we add a margin term M to constrain the value of $\mathcal{L}(s_-, o)$:

$$\mathcal{L}_{CN} = \mathcal{L}(s_+, o) + \alpha * \max\Big(M - \mathcal{L}(s_-, o), 0\Big)$$

For easy negatives with $\mathcal{L}(s_-, o) > M$, their effects won't be taken into account during training as the model can confidently distinguish them from positive samples.

Constrained Contrast We augment the ordinary loss function for summarization with a constrained contrastive term:

$$\mathcal{L}_{CC} = \mathcal{L}(s_+, o) + \alpha * \max\Big(\mathcal{L}(s_+, o) + M - \mathcal{L}(s_-, o), 0\Big)$$

In this formula, the model is not only trained towards predicting correct labels but also deviating from certain factual errors extracted from the contrast between the negative and positive samples.

4 Experiments

4.1 Dataset

We collected 10,000 samples of negative customer feedback from the post-order communication channels of e-commerce stores. We asked subject matter experts to generate summary for each customer feedback text with emphasis on extracting the information related to product issues. The summary is required to contain the (1) primary item names and (2) descriptions about the product defects associated with the items, if they are presented in the customer feedback. We use the human-produced summary as the target summary in model training. The train/test split ratio is 85:15.

Model	ROUGE-1	ROUGE-2	ROUGE-L
BART$_{+corruption,\mathcal{L}_{DC}}$	+0.30	+0.36	+0.49
BART$_{+corruption,\mathcal{L}_{CN}}$	+0.54	+0.01	+0.59
BART$_{+corruption,\mathcal{L}_{CC}}$	**+0.83**	**+1.12**	**+0.68**
T5$_{+corruption,\mathcal{L}_{DC}}$	+0.05	-0.19	+0.04
T5$_{+corruption,\mathcal{L}_{CN}}$	+0.20	+0.08	+0.25
T5$_{+corruption,\mathcal{L}_{CC}}$	**+0.45**	**+0.71**	**+0.43**

Table 3: Impact of our approaches on ROUGE scores. The reported numbers are relative changes of ROUGE scores compared to the ordinary fine-tuned BART and T5 models, respectively[4].

4.2 Model

We use two recently proposed abstractive summarization models, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), for customer-feedback summarization. We adopt the pretrained models from the HuggingFace implementation [3] and fine tune the models on our training dataset. Both models share the same training parameters including learning rate as 5e-5, $\alpha = 0.05$ in $\mathcal{L}_{DC}$, $(\alpha = 0.5, M = 2.0)$ in $\mathcal{L}_{CN}$, and $(\alpha = 0.5, M = 5.0)$ in $\mathcal{L}_{CC}$.

4.3 Evaluation metrics

We employ the ROUGE-1, ROUGE-2, and ROUGE-L scores (Lin, 2004) to ensure that our proposed methods do not degrade the fluency and continuity of the generated summary. These ROUGE scores measure the accuracy based on unigrams, bigrams, and longest subsequences.

We rely on the human evaluation to examine the factual consistency of the model output. We ask human annotators to classify the faithfulness of generated summary into *consistent* and *inconsistent* based on whether there are inaccurate or contradictory facts. We then compare the summary consistency before and after implementing the proposed methods.

5 Results

5.1 ROUGE Scores

We report the changes of ROUGE scores[4] in Table 3. Results show that the models trained with our correction methods generally have improvements on the ROUGE scores compared to the original BART and T5 models. Higher scores imply that the summaries from the corrected models are better aligned with the target summaries. In addition, using $\mathcal{L}_{CC}$ as the loss function turns out to produce the highest ROUGE scores for both BART and T5. Thus, for human evaluation, we will focus on the summaries produced by the models trained with $\mathcal{L}_{CC}$.

[3] https://huggingface.co/transformers/

[4] Absolute ROUGE scores are not shown due to confidentiality.

Model	Error Type	% Corrected
BART	WED	63.6
	IPD	50.0
T5	WED	46.7
	IPD	42.1

Table 4: Percentage of corrected WED and IPD errors for BART and T5. Comparisons are made between the ordinary models and the models trained with $\mathcal{L}_{CC}$.

Model	% Consis. to Inconsist.
BART	1.2
T5	2.1

Table 5: Percentage of cases where the summaries from the ordinary models are factual consistent but become inconsistent after our methods are applied.

5.2 Human Evaluation and Analysis

The human evaluation included 124 examples for BART and 600 examples for T5, all of which were randomly sampled from the test set. Table 4 shows the effect of our approaches on correcting the two major factual errors. As the results show, a large portion of the WED and IPD errors are corrected. Over 63% WED and 50% IPD mistakes from ordinary BART are rectified. For T5, our methods are able to correct around 46% WED and 42% IPD errors. It implies our models perform more robustly on the cases that can potentially lead to WED and IPD.

One remaining question is whether our approaches would degrade the originally faithful summaries. In Table 5, we report the percentage of cases where the summaries from the ordinary mod-

Source: (...) I bought this expensive TV that's supposed to have good screen and built-in wifi connection. But this one runs with lots of lagging, not as advertised on the website. (. . .) **Original**: Screen runs with lots of lagging, not as advertised. **After**: TV runs with lots of lagging, not as advertised.
Source: (. . .) The packaging is heavily damaged and opened, though the product inside is not broken. The seller should be careful on the packaging next time (. . .) **Original**: The packaging is heavily damaged and opened. Product is broken. **After**: The packaging is heavily damaged and opened. The product inside is not broken.

Table 6: Examples of error corrections using our methods.

els are consistent but become inconsistent after using our methods. We can see that most of the summaries remain consistent from our models. Furthermore, our analysis shows that the overall amounts of inconsistent summaries are reduced by 44.1% for BART and 31.6% for T5, which indicates the effectiveness of our methods.

Table 6 shows several input texts and summaries from the models before and after using our methods. In the first example, our model is able to pick up the correct entity from multiple entities in the source document, where the ordinary model fails. In the second example, the summary from the ordinary model contains contradicting description against the source document but our model captures the correct information.

6 Conclusion

In conclusion, we study the error patterns in the customer-feedback summaries generated by BART and T5. We propose to augment the training data with artificially corrupted summaries and use contrastive learning methods to enhance the model faithfulness. Human analysis shows that significant portion of WED and IPD errors from BART and T5 are reduced. Because our methods do not involve modifying the model structure, they can also be applied to other abstractive summarization frameworks.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR.

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.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

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.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

Tobias Falke, Leonardo FR 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.

Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2020. From standard summarization to new tasks and beyond: Summarization with manifold information. *arXiv preprint arXiv:2005.04684*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 540–551.

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.

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.

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.

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

Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained abstractive summarization: Preserving factual consistency with constrained generation. *arXiv preprint arXiv:2010.12723*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gucehre, 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

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.

Peng Yuan, Haoran Li, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. On the faithfulness for e-commerce product summarization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5712–5717.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. *North American Chapter of the Association for Computational Linguistics (NAACL) 2021*.