Prevent Mistakes, but Retain World Knowledge: Factual Summarization by Constraining Hallucinated Entities

Anonymous EMNLP submission

Abstract

State-of-the-art summarization models produce summaries that are fluent, but often factually incorrect. Many recent works attempt to mitigate this issue by making summaries faithful to the source document, at the expense of including less relevant out-of-source world knowledge. We seek to improve the factuality of model generated summaries while retaining factual extrinsic hallucinations. We propose an inference-time algorithm that integrates an entity-level factuality classifier to iteratively prevent the generation of non-factual entities. Applying our method to both BART and PEGASUS results in improvements in summary factuality. For BART, relative to a baseline of 42%, we show that our method improves factuality to 53%. Our method is competitive with other approaches for factual abstractive summarization without requiring fine-tuning.

1 Introduction

Despite tremendous improvement in abstractive summarization models to provide fluent summaries, recent benchmarks show that up to 60% of generated summaries contain factually incorrect statements (Pagnoni et al., 2021). On summarization tasks for highly abstractive datasets like XSum (Narayan et al., 2018), previous work shows that about 2 in 3 factuality errors correspond with *extrinsic hallucinations*: language in a summary that is not directly supported by the source article (Maynez et al., 2020).

There are differing opinions in the literature on whether *factual* extrinsic hallucinations are undesirable; even if those hallucinations are factual (correct given the source document and relevant world knowledge). Many papers propose approaches with the direct goal of reducing all extrinsic hallucinations (Nan et al., 2021; Chen et al., 2021). Considering that the vast majority of XSum ground truth summaries contain extrinsic hallucinations

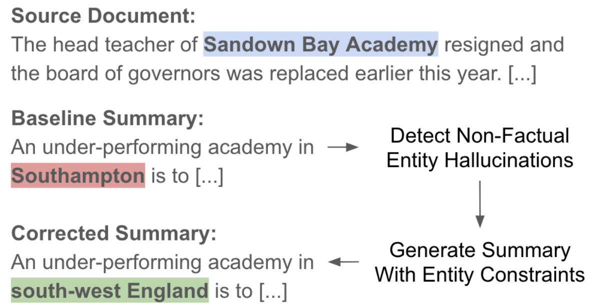

Figure 1: Example correction. Neither of the highlighted entities are present in the source. The baseline generated summary incorrectly states that *Sandown Bay Academy* is in *Southampton*, whereas GEF corrects the location to *south-west England*.

(Maynez et al., 2020), this work asserts that extrinsic hallucinations are not undesirable as long as they are factual. We seek to improve summary factuality by targeting non-factual extrinsic hallucinations without systematically excluding all extrinsic hallucinations.

We take inspiration from recent works by Cao et al. (2022) on detecting factual entity hallucinations using entity generation probabilities and King et al. (2022) who constrain beam search for summarization. We combine these research directions to propose an inference-time algorithm: **G**eneration via **E**ntity **F**actuality (GEF): an approach to iteratively detect and constrain the generation of non-factual entities to improve the factuality of generated summaries[1]. See Figure 1 for a high-level overview of GEF.

We demonstrate the utility of our method on XSum by evaluating the factuality of summaries generated by GEF compared to summaries generated by other summarization models in a human evaluation study ($N = 100$ documents). GEF improves upon BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020) baselines by 11% and

[1]Our code is available at `https://anonymous.4open.science/r/GEF-B414/`

7% respectively in terms of summary factuality, is competitive with other approaches for factual abstractive summarization (King et al., 2022; Cao et al., 2022) without requiring any fine-tuning, and performs best in terms of retaining factual extrinsic entity hallucinations (world knowledge). We analyze the semantic changes introduced by our method and find that most improvements stem from correcting dates, numbers and geographical locations.

2 Related Work

Summary Faithfulness and Factuality A recent large scale evaluation ($N = 500$) of abstractive summarization models finds that upwards of 60% of summaries generated by a BERT model trained on XSum contain extrinsic hallucinations. Moreover, over 70% of these generated summaries contain factuality errors. The study also introduces the intrinsic/extrinsic hallucination and factuality language that we leverage throughout our work (Maynez et al., 2020). Pagnoni et al. (2021) run a smaller evaluation ($N = 250$) on XSum and find that 83% of BERT-generated summaries contain factuality errors.

Efforts to improve the *faithfulness* of generated summaries aim to minimize extrinsic hallucinations. Both Chen et al. (2021) and Cao et al. (2020) artificially corrupt summaries with entity-swapping to learn how extrinsic entity hallucinations can be replaced with entities from the source. Nan et al. (2021) filter the XSum dataset to examples that do not contain extrinsic hallucinations and build a model which jointly learns how to detect summary-worthy entities during summarization. Narayan et al. (2021) prompts a transformer decoder with ordered sequences of entities extracted from the source when generating summaries. In contrast to these works which aim to improve faithfulness and remove extrinsic hallucinations, we seek to retain factual extrinsic hallucinations.

King et al. (2022) constrain beam search to improve the faithfulness of generated summaries based on a static set of rules. We propose and use a similar decoding technique integrated with a learned classifier.

Factuality Classification Filippova (2020) first suggest comparing the probabilities of an unconditional and conditional language model to inform the faithfulness and factuality of generated summaries for a data-to-text generation task. Cao et al. (2022) build upon that work by training an entity-level factuality classifier, which is used to facilitate the training of a factuality-aware summarization model. Our work is most similar to this line of work. While Cao et al. (2022) use their entity-level factuality classifier to fine-tune a BART model for summarization using reinforcement learning, our work does not require any fine-tuning; GEF instead uses this classifier to prevent the generation of non-factual entities during decoding.

3 Method

Let x be a source document and $\mathcal{Y}$ be the set of possible summaries for the source document. Define $\text{ENT}(y)$ as the set of entity spans in the summary, and y_e to be the entity mention for any $e \in \text{ENT}(y)$. The goal of GEF is to generate a summary, $y \in \mathcal{Y}$, such that all named entities in the generated summary are factual. Assuming access to a summarization model, $P(y \mid x)$, and an entity factuality classifier, FACT(x, y, e) (1 if factual 0 otherwise), we seek to optimize,

$$\begin{aligned} \max_{y \in \mathcal{Y}} \quad & P(y \mid x) \\ \text{s.t.} \quad & \text{FACT}(x, y, e) = 1 \text{ for all } e \in \text{ENT}(y) \end{aligned}$$

3.1 Detecting Non-Factual Entity Hallucinations

For the FACT constraint, we build upon Cao et al. (2022) who develop a classifier for detecting non-factual entity hallucinations. The classifier is a k-Nearest Neighbors binary classification model. It classifies whether each entity $e \in \text{ENT}(y)$ in a generated summary is factual by determining the majority class of the closest neighbors ($k = 20$) in the feature space. The model has two features: (1) the prior probability, ϕ_{prior}, of generating the entity *unconditioned* on the source document, but conditioned on the rest of generated summary, y_{-e}, and (2) the posterior probability, $\phi_{posterior}$, of generating the entity conditioned on the source document and the rest of the generated summary. We use a masked language model to compute the first term and a conditional masked language model to compute the second.

$$\phi_{prior} = \text{P}_{mlm}(y_e \mid y_{-e}) \tag{1}$$

$$\phi_{posterior} = \text{P}_{cmlm}(y_e \mid y_{-e}, x) \tag{2}$$

For P_{mlm} we use a non-fine-tuned BART-Large. For P_{cmlm} we use the BART-Large fine-tuned by

Algorithm 1 Generation via Entity Factuality

$x \leftarrow$ input document
$\mathrm{EX} \leftarrow \emptyset$
for $i = 1$ to ... **do**
  $y^{(i)} \leftarrow \mathrm{beamsearch}(x, \mathrm{EX})$
  $\mathrm{ENT} \leftarrow \mathrm{NER}(y^{(i)})$
  $\mathrm{EX}' \leftarrow \{e \in \mathrm{ENT} : \mathrm{FACT}(x, y^{(i)}, e) = 0\}$
  if $\mathrm{EX}' = \emptyset$ **then**
    return $y^{(i)}$
  $\mathrm{EX} \leftarrow \mathrm{EX} \cup \mathrm{EX}'$

Cao et al. (2022) for masked language modeling on XSum, conditional on the source document.

3.2 Generation via Entity Factuality (GEF)

To approximate our objective function, we utilize an iterative constrained beam search approach shown in Algorithm 1. In the first iteration ($\mathrm{EX} = \emptyset$) we perform standard beam search decoding to produce a hypothesis summary $y^{(1)}$. Next, we add non-factual entities to the blacklist set EX. Subsequent iterations of GEF uses EX to decide which entities it cannot produce in attempt to generate factual summaries.

Beam search is run again with generated phrases matched against phrases in EX. If a phrase matches the blacklist, the partial generation is removed from the beam. Compound entities are split into sub-entities to enable granular detection of non-factual parts. The algorithm terminates when all entities in the generated output are deemed to be factual or when every beam hypothesis is non-factual.

4 Dataset and Evaluation

We evaluate our method on a subset of XSum Test ($N = 10,875$). This subset excludes the training data for the factuality classifier previously annotated by Cao et al. (2022) ($N = 459$). We do not define a validation set since GEF is not fine-tuned.

Measuring Factuality We consider a generated summary y to be factual if all of its detected entities $\mathrm{ENT}(y)$ are factual or if the summary contains no entities. This assumption is motivated by the fact that most factual errors stem from entities (Pagnoni et al., 2021; Nan et al., 2021).

For evaluating factuality we use generated summaries in XSum Test where the baseline models contain at least one extrinsic entity hallucination (60.2% for BART and 61.7% for PEGASUS). From these sets, we randomly sample 100 documents and annotate summaries generated by every model. We follow Cao et al. (2022)'s approach for labeling the factuality of entities. An entity is extrinsically hallucinated if it is not contained within the source. Annotators perform an online search to determine whether entity hallucination factually represent world knowledge. An entity contained within the source is an intrinsic hallucination if it misrepresents the source document, and non-hallucinated otherwise. All intrinsic hallucinations are considered to be non-factual. In total we label 2,661 entities across 730 unique summaries.[2]

Model Comparisons We run GEF correction on top of BART-Large and PEGASUS fine-tuned on XSum (Lewis et al., 2020; Zhang et al., 2020).

We compare with other BART-based approaches for factual abstractive summarization. Our closest comparison is a reinforcement learning model with $r_{\mathrm{nfe}} = -2.0$ (RL-Fact) by Cao et al. (2022) which seeks to retain factual extrinsic hallucinations and leverages the same entity classifier. We also consider two approaches which aim to improve factuality by optimizing for summary faithfulness: Entity Corrector (Chen et al., 2021) and PINOCCHIO (King et al., 2022).

To assess upper-bound performance of this approach we evaluate GEF Oracle. The oracle leverages human annotations to determine factuality of entities within a summary. This evaluates performance with an entity classifier that achieves 100% accuracy in detecting non-factual hallucinations.

5 Main Results

Table 1 shows the evaluation results. GEF outperforms the BART-Large baseline by a significant margin: 11% more generated summaries are factual with a slight decrease (-0.48) in ROUGE-1 (Lin, 2004). The model also retains the same amount of factual extrinsic hallucinations as the base model. RL-Fact produces slightly more factual summaries, but has 5% fewer summaries with factual extrinsic entity hallucinations than the baseline and a bigger drop in ROUGE-1 (-1.20). Both PINOCCHIO and Entity Corrector do not improve significantly upon the baseline in terms of factuality; however these methods are mainly designed to target faithfulness as opposed to factuality. For PEGASUS,

[2]Table 3 in the appendix shows the distribution of the labeled entities. The annotations are published online: https://anonymous.4open.science/r/GEF-B414/data/xsum/gold-metrics.json.

Base System	Model	Factual ↑ Summaries	Sum. w/ Factual ↑ Extrinsic Entities	ROUGE R1	ROUGE R2	ROUGE RL
BART-Large	Baseline	42%	44%	45.23	22.18	37.02
	GEF	**53%**	44%	44.75	21.57	36.46
	RL-Fact	**55%**	39%	44.03	21.18	36.02
	Entity Corrector	42%	45%	44.84	21.70	36.71
	PINOCCHIO	43%	43%	44.37	21.22	36.08
	GEF Oracle	67%	63%	-	-	-
	Ground Truth	100%	88%	-	-	-
PEGASUS	Baseline	58%	61%	46.73	24.37	38.95
	GEF	65%	54%	46.29	23.43	38.30
	GEF Oracle	79%	74%	-	-	-
	Ground Truth	100%	93%	-	-	-
# of samples		$N = 100$		$N = 10,875$		

Table 1: Evaluation on XSum Test with BART-Large and PEGASUS. Factuality evaluation is based on our annotations for 100 summaries per system (see Section 4). A summary is factual if all of its entities are annotated as factual. "Sum. w/ Factual Extrinsic Entities" is the percentage of summaries that have at least one extrinsic entity, and all of the extrinsic entities are factual. We do not compute ROUGE scores for GEF Oracle since it depends on fully annotated data. Statistically significant (Wilcoxon, $p < 0.01$) improvements upon the baseline are bolded.

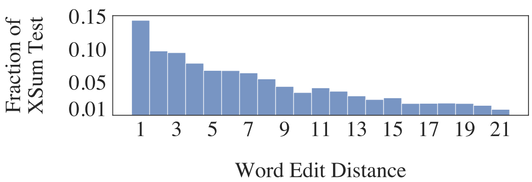

Figure 2: Edit distance on GEF BART corrected summaries for XSum Test.

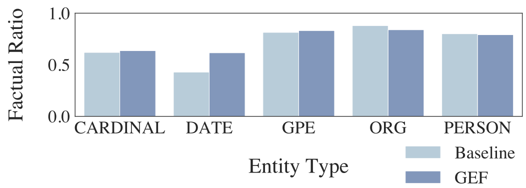

Figure 3: Changes in factual entity types between BART baseline and GEF corrected summaries for the top 5 entity types in the human annotated dataset ($N = 100$).

we find that GEF improves upon the baseline by 7%. Note, for these experiments GEF is identical to the BART experiments, and does not require retraining. The performance of GEF with an oracle shows that there is room for improvement through the entity factuality classifier.

What Changes Does GEF Make? Figure 2 shows word edit distance between the baseline and GEF on XSum Test. A majority of summaries (68.5%) remain unchanged as baseline summaries contain no extrinsic hallucinations (60.2%), and the classifier predicts that the baseline summaries are factual (8.3%). Most of GEF's corrections (53.8%) involve changing at most 6 words. Figure 3 shows how GEF's corrections are distributed among the top 5 entity types in the human-annotated dataset. Most of GEF's factuality improvements stem from correcting more general entity classes such as dates, numbers, and geographical locations.

Efficiency The average number of GEF iterations to complete a summary is 2.84 for BART and 2.78 for PEGASUS for all of XSum Test. The majority of summaries complete by 3 iterations, 85.6% for BART and 87.6% for PEGASUS.

6 Conclusion

This work shows that integrating a factuality classifier into inference can significantly improve the factuality mistakes of a system without eliminating world knowledge or requiring fine-tuning. Future methods could improve further by increasing the accuracy of the factuality classifier, as evidenced by the relatively high oracle score.

Limitations

Our approach is fundamentally limited by the limits of the fine-tuned summarization model since we only make corrections at inference time. Further, it might be computationally prohibitive in low-resource settings since it requires fine-tuning one model for summarization and another for computing $\phi_{posterior}$, and running a third pre-trained model for computing ϕ_{prior}. We focus on correcting extrinsic entity hallucinations, whereas a significant amount of factuality errors stem from intrinsic hallucinations.

Ethical Implications and Broader Impact

Using language models pre-trained on massive sets of unfiltered data for summarization introduces biases. This work attempts to mitigate this issue by correcting factuality errors in generated summaries, however we do not target other sources of bias and harm possibly introduced by pre-training on these massive datasets. If summarization systems become sufficiently factual it might lead to less human auditing of model output, however it is important to assess other qualities of automated summarization besides correctness to ensure that they have a positive societal impact.

References

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

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.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. 2022. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. *CoRR*, abs/2203.08436.

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.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. 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.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

A Annotation Process

The authors of this paper annotated factuality of entities in the generated summaries following the same labeling approach as Cao et al. (2022). The distribution of labels is shown in Table 3 and the annotation workflow is available at this url: `https://anonymous.4open.science/r/GEF-B414/annotation_demo.ipynb`.

B Implementation Details

We use Hugging Face's transformers (Wolf et al., 2020), PyTorch for model inference (Paszke et al., 2019) and spaCy[3] for named entity recognition.

Model Hyperparameters We use default beam search hyperparameters to run GEF on BART and PEGASUS: min length = 11, 0; max length = 62, 64; length penalty = 1.0 (no penalty), 0.6. We fix the number of beams to 4 for both GEF BART and PEGASUS for consistency.

To ensure termination, we set a large upper bound (100) for the number of iterations. Empirically, we find that all summaries complete after 11 iterations for BART and 12 for PEGASUS.

Computing Infrastructure Generating summaries with GEF for all of XSum Test takes about 2 hours with a batch size 16 on a single RTX5000 GPU.

Model Checkpoints For GEF's generative models, we use the published Hugging Face checkpoints for BART-Large[4] and PEGASUS[5] fine-tuned on XSum. The checkpoint used to compute ϕ_{prior} is also published on Hugging Face [6]. We use the model checkpoint and data published[7] by Cao et al. (2022) to compute $\phi_{posterior}$ and train their k-Nearest Neighbor factuality classifier.

Datasets and Evaluation All the code to reproduce our experiments is available in our code repository.[8]. The XSum dataset is available in Hugging Face's dataset directory[9].

C Model Output

Table 2 shows examples of two typical corrections made by GEF. In the first example, BART incorrectly hallucinates the location of the Minions world premiere to be in Los Angeles. GEF corrects the location of the event to London without making significant changes to the summary. Notably, London is a factual extrinsic hallucination not contained within the source. In the second and third examples, BART adds an incorrect detail to the generated summaries, which is removed by GEF in the corrected summaries.

[3] `https://spacy.io/`

[4] `https://huggingface.co/facebook/bart-large-xsum`

[5] `https://huggingface.co/google/pegasus-xsum`

[6] `https://huggingface.co/facebook/bart-large`

[7] `https://github.com/mcao516/EntFA`

[8] `https://anonymous.4open.science/r/GEF-B414/evaluate_summaries.py`

[9] `https://huggingface.co/datasets/xsum`

BART	The world premiere of Minions has taken place in Los Angeles, USA.
GEF	The world premiere of the new animated film Minions has taken place in London.
BART	The air is still as charged as it was when I first arrived in Ravenscraig 25 years ago.
GEF	The air is still as charged as it was when I first arrived in Ravenscraig, the site of the last British steelworks to close.
BART	Liam Payne and Niall Horan have confirmed they're staying in One Direction.
GEF	One Direction have confirmed they will continue together.

Table 2: Examples of factuality corrections made by GEF.

Model	Labeled Entity Hallucinations				
	Factual	**Non-factual**	**Intrinsic**	**Non-hallucinated**	**Total**
BART Baseline	90	72	13	186	361
BART GEF	78	64	17	194	353
BART GEF Oracle	104	15	38	187	344
Entity Corrector	85	66	19	188	358
Pinocchio	89	71	19	183	362
RL-Fact	77	52	20	154	303
Ground Truth	192	0	0	169	361
PEGASUS Baseline	123	47	9	198	377
PEGASUS GEF	103	39	14	186	342
PEGASUS GEF Oracle	134	17	9	184	344
Ground Truth	183	0	0	198	381

Table 3: Human-annotated entity labels for model-generated summaries in two sets of 100 sample documents. Ground truth summary entities are assumed to be factual.