Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Selection

Sihao Chen[1]* **Fan Zhang**[2] **Kazoo Sone**[2] **Dan Roth**[1]*
[1]University of Pennsylvania [2]Google
{sihaoc, danroth}@cis.upenn.edu, {zhanfan, sone}@google.com

Abstract

Despite significant progress in neural abstractive summarization, recent studies have shown that the current models are prone to generating summaries that are *unfaithful* to the original context. To address the issue, we study *contrast candidate generation* and *selection* as a model-agnostic post-processing technique to correct the extrinsic hallucinations (i.e. information not present in the source text) in unfaithful summaries. We learn a discriminative correction model by generating alternative candidate summaries where named entities and quantities in the generated summary are replaced with ones with compatible semantic types from the source document. This model is then used to select the best candidate as the final output summary. Our experiments and analysis across a number of neural summarization systems show that our proposed method is effective in identifying and correcting extrinsic hallucinations. We analyze the typical hallucination phenomenon by different types of neural summarization systems, in hope to provide insights for future work on the direction.

1 Introduction

Abstractive Summarization is the task of producing a concise and fluent summary that is salient and *faithful* to the source document(s). Data-driven, neural methods (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017), and the more recent, pretrained transformer language models (Vaswani et al., 2017; Devlin et al., 2019; Liu and Lapata, 2019), have shown improvements in the fluency and salience of generated summaries.

However, less progress has been made on improving the *faithfulness* of the generated summaries, that is, producing a summary that is *entailed* by the information presented in the source document. Despite the increased level of performance under automatic metrics such as ROUGE (Lin, 2004) or BERTSCORE (Zhang et al., 2020), current state of the art models (Liu and Lapata, 2019; Lewis et al., 2020) produce summaries that suffer from *intrinsic* and *extrinsic hallucinations* – the fabrication of untruthful text spans containing information either *present* or *absent* from the source (Maynez et al., 2020).

Source: He was re-elected for a second term by the UN General Assembly, unopposed and unanimously, on **21 June 2011**, with effect from 1 January 2012. Mr. Ban describes his priorities as mobilising world leaders to deal with climate change, economic upheaval, pandemics and increasing pressures involving food, energy and water...
Unfaithful Summary: The United Nations Secretary-General Ban Ki-moon was elected for a second term in **2007**.
Our Summary: The United Nations Secretary-General Ban Ki-moon was elected for a second term in **21 June 2011**.

Table 1: An example *unfaithful* summary. It suffers from *extrinsic* hallucination, where information not present in the source document was generated. Our method attempts to correct the unfaithful summary by replacing "2007" with an entity from the source with compatible semantic type (i.e. DATE).

Table 1 shows an example of such summary, generated by BART (Lewis et al., 2020), an auto-regressive, transformer-based sequence-to-sequence model. The article describes an event where the former UN-Secretary-General Ban Ki-Moon was re-elected for a second term. The model hallucinates "2007", which never appears in the source document, leading to inconsistency with the correct date of the event presented.

In this work, we focus on the problem of *correcting* such hallucinations as a post processing step[1]. A post processing correction step allows us to rely on the fluency of the text generated by SOTA systems, that gain from huge pretrained models and large fine-tuning datasets, and correct it using small

* Most of the work done while the authors were at Google.

[1]Our code and data is available at http://cogcomp.org/page/publication_view/938

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5935–5941
June 6–11, 2021.

amounts of automatically generated training data.

Under the setting where a large fraction of ground truth summarization data is hallucinated, as we show in Table 2, we study the method of *contrast candidate generation* and *selection*. In the generation step, we replace named entities in a potentially hallucinated summary with ones with compatible semantic types that are present in the source, and create variants of candidate summaries. In the selection step, we rank the generated candidates with a discriminative model trained to distinguish between faithful summaries and synthetic negative candidates generated given the source. We experiment on a range of RNN- and transformer-based abstractive summarization models. Our preliminary results on the XSum corpus (Narayan et al., 2018a), which contains substantial presence of hallucinated ground truth examples, show the effectiveness of our method in correcting unfaithful summaries with *extrinsic hallucinations*.

Our main contributions are as follows. First, our work is the first to study the effectiveness of *contrast candidate generation* and *selection* as a model-agnostic method for correcting hallucinations, under the setting where a large fraction of ground truth summarization data suffers from hallucinations. Second, we validate our method on various neural summarization systems trained on XSum, and provide detailed analysis on the typical types of hallucinations from each system.

2 Contrast Candidate Generation & Selection

Our proposed method is built on the observation that a large fraction of extrinsic hallucinations happen on named entities and quantities. Table 2 shows the human analysis by Maynez et al. (2020) on the hallucinations of 500 randomly sampled gold summaries from the XSum corpus . We break down each category and annotate the proportion of hallucinations that happen on entity and number/quantity spans.

As Maynez et al. (2020) further show that the hallucinations in training data translate to similar issues for the generated outputs across different summarization models, we want to study a model-agnostic, post-processing method that can correct such entity and quantity hallucinations. We frame the problem as a correction task and make it conceptually a less complex problem than summarization. Modeling correction as a standalone task would require less training data, which becomes crucial when a large proportion of ground truth summarization data suffer from hallucinations, and inherit the fluency of data intensive SOTA models.

Type	%	Ent. %	Num. %
Faithful	23.1	-	-
Ex. Hallucination	73.1	35.9	18.2
In. Hallucination	7.4	1.9	0.5

Table 2: Frequency of extrinsic and intrinsic hallucinations in 500 ground truth summary of the XSum corpus. The "%" column shows the % of intrinsic and extrinsic hallucinations annotated by Maynez et al. (2020). We analyzed the % of hallucinations on entities and numbers/quantities, and show the % out of all 500 summaries in the right two columns.

2.1 Contrast Candidate Generation

From a model-generated summary, we first identify any potentially hallucinated entities or quantities by checking whether entities with similar surface forms have appeared in the source document. We use a neural Named Entity Recognition (NER) system from the Stanza NLP toolkit (Qi et al., 2020) trained on the OntoNotes corpus (Weischedel et al., 2013) to extract named entities of different semantic types from the source document and summary. Each named entity present in the summary is replaced with a different entity present in the document with the same NER label. This gives us different variants of the original summary with the same level of fluency , but not necessarily faithful.

2.2 Contrast Candidate Selection

For the candidate selection step, we want to identify the best candidate among the variants generated in the previous step as the final output summary. As the contrast candidates vary in no more than a few tokens from the original summary, it requires a model with more delicate local decision boundaries (Gardner et al., 2020) to select the correct candidate. For example, we observe that MNLI models (Williams et al., 2018) fail to produce satisfactory results.

To create training data for that purpose, we sample examples from the XSum training set where all entities in the ground truth summary appear in the source document. We then follow the same procedure in the generation step, and produce *unfaithful* variants from the ground truth summary by replacing entities with others that have the same se-

mantic type but different surface form in the source text. With the ground truth and synthetic negative summaries, we train a text classifier with a discriminative objective to score and rank the variants of the summaries.

We use BART (Lewis et al., 2020) plus a linear layer as our classification model. We adopt a similar learning objective to contrastive learning (Khosla et al., 2020). For each pair of positive and negative summary candidate, we use cross entropy loss $\mathcal{L}_{\text{XE}}$ to handle the correctness of the label predictions. We add a margin ranking loss term $\mathcal{L}_{\text{RANK}}$ to encourage the model to assign higher probability to the positive than the negative candidate. The margin γ is a tunable hyperparameter in training.

$$\mathcal{L} = \mathcal{L}_{\text{XE}}(\hat{y_+}, 1) + \mathcal{L}_{\text{XE}}(\hat{y_-}, 0) + \mathcal{L}_{\text{RANK}}(\hat{y_+}, \hat{y_-})$$
$$\mathcal{L}_{\text{RANK}} = max(0, \hat{y_-} - \hat{y_+} + \gamma)$$

During test time, we use the trained model to score the generated contrast candidate summaries, as well as the original version generated by the summarization model. We take the candidate with the highest score as the final summary.

3 Experiments

Full XSum Test Set			
Method	ROUGE$_L$	BERT	FEQA (%)
BART$_{large}$	**36.95**	**91.57**	-
+ correct	36.70	91.50	-
Changed Summary Only (13.3%)			
BART$_{large}$	**38.63**	**91.61**	22.50
+ correct	36.62	91.10	**25.62**

Table 3: Evaluation with automatic metrics on the summaries generated by the baseline BART$_{large}$ model, plus our post-processing correction method. We report $F_{\beta=1}$ scores with ROUGE and BERTSCORE, plus the macro-averaged percentage of questions answered correctly for each summary with FEQA, a QA-based metric for summary faithfulness proposed by Durmus et al. (2020).

Our experiments focus on the aforementioned XSum corpus, where the target summary is highly *abstractive* and likely *hallucinated*. We first consider the summaries generated by a BART model trained on the XSum corpus. By applying our method, we are able to change 13.3% of all model generated summaries. For 38.4% of all summaries, the original summary does not have a hallucinated entity, or there is no entity with compatible type in the source text. Our model decides to keep the original summary in the rest 48.3%.

Method	*Faith. %*	*Ex. %*	*In. %*
BART	23.8±9.6	71.7±11.2	**1.7**±3.5
+ correct	**59.5**±12.4	**9.2**±7.3	29.1±11.6

Table 4: Percentage of examples human annotator judged as "faithful" (*Faith.*), "extrinsically hallucinated" (*Ex.*), and "intrinsically hallucinated" (*In.*) among the 95 randomly sampled corrected summaries. The 95% confidence intervals are estimated with bootstrap resampling (Appendix. C).

3.1 ROUGE and BERTSCORE Evaluation

We first verify that our method does not hurt the fluency and salience of the generated summaries, for which we assume ROUGE (Lin, 2004) and BERTSCORE (Zhang et al., 2020) are suitable metrics. We report the results in Table 3. We observe though both the baseline and our method do well in both ROUGE and BERTSCORE, our method trails behind in both metrics slightly. This is due to the existence of extrinsic hallucinations in the ground truth summary, and the model manages to generate a part of the hallucinations, and gets incorrectly rewarded by such.

3.2 Faithfulness Evaluation

To test whether our correction method can improves the faithfulness of the summaries, we evaluate the summaries with FEQA (Durmus et al., 2020), a QA-based metric for summary faithfulness. Given a summary, FEQA automatically generates questions on noun phrase and named entity spans in the summary, and uses a pretrained QA model to verify if the answer derived from the source document exact-matches the span in the summary.

We run FEQA and compute the macro-averaged percentage of questions answered correctly for each of the 1510 summaries that our system made corrections to, and report the results in Table 3. The results suggest that the corrected summaries present statistically significant improvements over the original ones ($p < 0.001$, with a two-tailed, paired t-test).

Table 4 shows the human evaluation results on the 95 randomly sampled subset of changed summaries. Two expert annotators assign each summary into three faithfulness categories and adjudicate the decisions. Additional annotations from

Good Corrections		
Type	System	Original Summary and Our Change
Correcting NE Hallucination	BERTS2S	Tranmere Rovers have signed midfielder [Alfreton]$_{\text{PER}}$ → [Mooney]$_{\text{PER}}$ on loan until the end of the season.
Correcting Number Hallucination	BART	A judge has ruled that the [$9.6bn (£5.03bn)]$_{\text{MONEY}}$ → [$7.8bn (£5.03bn)]$_{\text{MONEY}}$ oil spill compensation fund is not fraudulent.
Typical Mistakes		
No correct replacement exists in source	BART	Helmut Kohl, who has died at the age of [87]$_{\text{CARDINAL}}$ → [39]$_{\text{CARDINAL}}$, was one of the driving forces behind Germany's reunification in 1990.
Wrong type of NE in summary	TRANS2S	[Andrew Marr]$_{\text{PER}}$ → [Venter]$_{\text{PER}}$ is one of the most important scientific discoveries in human life.
Not explicit in source, but can be inferred	BERTS2S	[Three]$_{\text{CARDINAL}}$ → [Two]$_{\text{CARDINAL}}$ fugitives have been arrested and charged with attempting to smuggle drugs into the country.

Table 5: Examples of corrections and typical mistakes made by our proposed method on generated summaries by different summarization models. The original and replaced entities in each summary are highlighted, and are colored by their faithfulness categories (Red: *Extrinsic Hallucation*; Orange: *Intrinsic Hallucation*; Blue: *Faithful*)

System	P	R	F_1	ENT. %
PTGEN	79.86	58.38	67.45	65.48
TCONVS2S	87.76	61.87	72.57	64.27
TRANS2S	81.81	57.35	67.44	80.15
BERTS2S	80.54	37.82	51.47	56.85

Table 6: The selection model's precision, recall and F_1 on identifying hallucinated output from four different summarization systems. The ENT. % column shows the % of hallucinations on entities and quantities among all hallucinated summaries by each system.

a third expert is then used to calculate the inter-annotator agreement. As the results show, our model is able to improve the faithfulness of the summaries, but at the cost of incurring intrinsic hallucinations on mistakes, which we will discuss more in detail in section 4.2.

4 Analysis and Discussion

4.1 Identifying Hallucination Across Systems

Table 6 shows our selection model's performance when measuring P, R, F_1 w.r.t all the hallucinated instances. We use the test set from Maynez et al. (2020), who have annotated hallucination categories of generated summaries from four neural summariazaiton models: PTGEN (See et al., 2017) TCONVS2S (Narayan et al., 2018a), BERTS2S and TRANS2S (Rothe et al., 2020). Our system achieves consistently high level of precision across models. The system achieves high relative recall with respect to the % of entity and quantity hallucinations among all hallucinations. As our method only targets entities and quantities, the overall recall varies by the typical type of hallucinations each summarization system makes. We also observe while our method achieves high recall on models with lower ROUGE and BERTSCORE, the recall drops on pretrained models such as BERTS2S. This is potentially due to the decreased percentage of entity/quantity hallucinations exist in generated summaries from the models with pretraining.

4.2 Intrinsic vs. Extrinsic Hallucinations Trade-off

As our method detects and corrects extrinsic-hallucinated entities, naturally any entities replaced wrong would introduce intrinsic hallucinations in the changed summary, as indicated by the results in Table 4. To speculate why the mistakes happen, we analyzed the typical mistakes by the model, and listed a few representative examples in Table 5. For example, our method could not find the correct replacement for a hallucinated entity when no such one exists in the source text. We observe that the models with pretraining, such as BERTS2S, (Rothe et al., 2020) and BART, suffer from the issue by most, as they tend to be affected by artifacts/priors from the pretraining process.

4.3 Entity Faithfulness $\subsetneq$ Summary Faithfulness

From the observation that models often hallucinate entities with no correct replacement in the source, we suspect that solving entity faithfulness alone does not guarantee the faithfulness of the summary. In the last example from Table 5, the BERTS2S system correctly identifies that three fugitives are involved in the event described by the source text, even though the number "three" has never been explicitly mentioned in the source context in any surface forms. Furthermore, statistics provided by Maynez et al. (2020) show that abstractive summarization models often produces *factual* statements,

i.e. verifiable in the real world independent of the source text. Such findings imply that identifying hallucinations often requires more complex objectives such as commonsense reasoning and knowledge retrieval. The solution we propose here that focuses only on entites and quantities would likely be insufficient to solve the entire problem.

5 Related Work

There have been growing interests in quantitatively measuring the faithfulness of text generation models. Most widely-adopted evaluation metrics for text generation, such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020), correlate poorly with the human perceived faithfulness of the generated text (Kryscinski et al., 2019; Durmus et al., 2020). Recent studies explore categorical, content-based analysis for measuring the faithfulness of summaries (Goyal and Durrett, 2020; Deutsch and Roth, 2020). Narayan et al. (2018b); Deutsch et al. (2020); Durmus et al. (2020) propose to use question answering to test the consistency of summary content to the information presented in the source text.

There have been efforts to study pre- or post- processing methods to improving faithfulness of generated summaries. Falke et al. (2019) attempt to use textual entailment models to re-rank the summary candidates generated from beam search or different neural systems. As Maynez et al. (2020) highlight the existence of hallucinations in training data, truncating potentially unfaithful gold summaries during training is an effective strategy (Kang and Hashimoto, 2020; Filippova, 2020). Kryscinski et al. (2020) take similar apporach as in this work to identify the hallucinations in summary. A concurrent study to this work (Cao et al., 2020) uses similar strategies as in this paper on a dataset with a very small fraction of hallucinations present. Our study instead focuses on the more challenging setting (Goyal and Durrett, 2021) where a large part of training data suffers from extrinsic and intrinsic hallucinations, and provides cross-system analysis on the both hallucinations categories.

6 Conclusion

We study contrast candidate generation and selection as a method to apply post-hoc fixes to extrinsically hallucinated summary on entities and quantities, under the setting where the summarization dataset suffers from intrinsic and extrinsic hallucinations. We conduct our experiments on the XSum dataset, and show that our method is able to correct extrinsic hullucinations, but incurs a small fraction of intrinsic hallucinations on mistakes. We also provide detailed analysis and discussions on the capabilities and limitations of our method. We hope our findings in the paper will provide insights to future work in this direction.

Acknowledgments

We thank Sunita Verma and Sugato Basu for valuable input and feedback on drafts of the paper. This work was supported in part by a Focused Award from Google, a gift from Tencent, and by Contract FA8750-19-2-1004 with the US Defense Advanced Research Projects Agency (DARPA). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

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.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2020. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *arXiv preprint arXiv:2010.00490*.

Daniel Deutsch and Dan Roth. 2020. Understanding the extent to which summarization evaluation metrics measure the information quality of summaries. *arXiv preprint arXiv:2010.12495*.

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.

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.

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.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Daniel Kang and Tatsunori Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Proceedings of the 34th Conference on Neural Information Processing Systems*.

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, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 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, 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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3721–3731.

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 GuÌ‡lçehre, 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. 2018a. 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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Candidate Selection Model

For our contrast candidate selection model, we use a pretrained BART base model. We add a linear layer over the max pooled embedding, and the classification model is expected to output a label between ["FAITHFUL", "HALLUCINATED"].
For all our experiments, we use the following set of hyper-parameters: $r = 1e-5$, margin $\gamma = 0$, number of training epoch$= 3$.

B Complete ROUGE and BERTSCORE Results

Full XSum Test Set				
Method	R_1	R_2	R_L	BERT
BART$_{large}$	45.10	21.86	36.95	91.57
+ correct	44.82	21.49	36.70	91.50
Changed Summary Only (13.3%)				
BART	46.73	23.51	38.63	91.61
+ correct	44.35	20.70	36.62	91.10

Table 7: ROUGE$_{\{1,2,L\}}$ and BERTSCORE evaluation results (in F_1) of summaries generated by the baseline BART$_{large}$ model, plus the corrected summaries with our post-processing method, on the test set of the XSum corpus.

C Estimating Confidence Interval for Human Evaluation

We use bootstrapping to estimate the confidence interval for the expert annotation presented in Table 4. For each faithfulness category on the two systems, we regard the adjudicated annotation as ground truth, and label the individual instance as the true positive (TP), false negative (FN), true negative (TN) and false positive (FP) respectively according the annotations from the third expert. We re-sample the 95 instances with replacement for 1,000 times. We estimate the adjusted mean and 95% confidence interval from the mean and standard deviation of the sampled distribution of (TP + FN).