Reducing named entity hallucination risk to ensure faithful summary generation

Eunice Akani[1,2] and **Benoit Favre**[1] and **Frederic Bechet**[1] and **Romain Gemignani**[2]
[1] Aix-Marseille Univ, CNRS, LIS, Marseille, France
[2] Enedis, Marseille, France
firstname.lastname@lis-lab.fr

Abstract

The faithfulness of abstractive text summarization at the named entities level is the focus of this study. We propose to add a new criterion to the summary selection method based on the "*risk*" of generating entities that do not belong to the source document. This method is based on the assumption that Out-Of-Document entities are more likely to be *hallucinations*. This assumption was verified by a manual annotation of the entities occurring in a set of generated summaries on the CNN/DM corpus. This study showed that only 29% of the entities outside the source document were inferrable by the annotators, leading to 71% of hallucinations among OOD entities. We test our selection method on the CNN/DM corpus and show that it significantly reduces the hallucination risk on named entities while maintaining competitive results with respect to automatic evaluation metrics like ROUGE.

1 Introduction

Abstractive text summarization methods aim at generating shorter versions of documents, possibly containing new words with respect to original documents. Recent pre-trained language models (Devlin et al., 2019; Lewis et al., 2020; Zhang et al., 2019; Raffel et al., 2020) allow to obtain fluent generated text. Despite their remarkable performance, those models tend to generate information that cannot be inferred from the source document. According to a study by Cao et al. (2018), 30% of the summaries generated by various systems have inconsistencies, qualified as "hallucination" by (Maynez et al., 2020). Current metrics used to assess automatic text summarization systems, such as ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020), do not account for these issues. These hallucinations led to several studies on evaluating the faithfulness of generated summaries and generating more faithful texts (Li et al., 2022; Ji et al., 2023). For example Durmus et al., 2020; Deutsch et al., 2021 proposed a QA-based metric. They produced questions from the generated summary and check that the answer in the document matches the one in the summary. Natural Language Inference (NLI) have also been proposed to evaluate the factuality of a generated summary by checking if it entailed the source document (Falke et al., 2019; Maynez et al., 2020; Laban et al., 2022). Most recent studies (Luo et al., 2023) used ChatGPT[1] as NLI system to evaluate the factual inconsistency of generated summary. Chen et al., 2021 suggested using contrast candidate generation and selection as a post-processing method to avoid hallucination. They create candidate summaries and replace named entities with ones found in the document to summarize. Nan et al., 2021 introduced an entity-based metric to compare the entities in the gold summary to the generated summary. They also proposed a method based on training data filtering and multi-task learning to avoid hallucinations.

This paper studies hallucinations at the entity level, which is a crucial level (Chen et al., 2021) in news-related corpora such as CNN/DM (Hermann et al., 2015) or XSum (Narayan et al., 2018). We aim to reduce the "*risk*" of generating nonfactual summaries by reducing the number of possibly hallucinated entities using a new criterion called *Named Entity Hallucination Risk* (*NEHR*). This criterion stands on the precision-source of Nan et al., 2021 for summary selection. Based on the assumption that an Out-Of-Document (OOD) entities have more chance to be hallucinations, we empirically verify that the entities in the summary are occurring in the source document.

Our contributions are as follows :

- We create summaries using sampling methods and we propose a selection criterion called *Named Entity Hallucination Risk* (NEHR) to minimize factual errors.

[1] https://openai.com/blog/chatgpt

- We empirically examine the relevance of NEHR as a selection criterion.
- We evaluate summaries chosen with NEHR and the model's performance on two benchmark corpora CNN/DM and XSum.

2 Named Entity Hallucination Risk

We propose to assess the risk of using entities and quantities incorrectly, leading to hallucinations in generated summaries. In the following, we call an *entity* a word string belonging to a named entity category such as person, location, organisation and quantities. An entity can be considered as an *in-document* entity if its corresponding word string can be found in the source document (*exact-match in-document entities)*, or if it relates to an alternative form (*e.g. New York/Big Apple*). If an entity is not mentioned in the source document, it is considered as an Out-Of-Document (OOD) entity. OOD entites can be either acceptable if they can be inferred from the document (*e.g. New York* → *USA*) or can be considered as *hallucinations*.

We call *hallucinations* the occurrence in generated summaries of entities that cannot be inferred from the source document. Hallucinations can be obvious errors, entities that have nothing to do with the source document context, or entities that could be acceptable by looking at external sources but which were considered as dubious by human annotators: accepting them would require some form of world knowledge beyond general understanding, not directly available from the document. Let's point out that an entity can be an in-document entity or an acceptable OOD entity but still be incorrect in its context of occurrence in a generated summary.

This study aims to reduce the risk of hallucinating entities thanks to a selection criterion called *NEHR* for *Named Entity hallucination Risk* inspired by the *precision-source* metrics proposed by Nan et al., 2021. A simple selection system can be built from our NEHR criterion, which selects less risky hypotheses from a large sample of summaries.

NEHR definition The NEHR criterion is based on the assumption that in-document entities, and moreover exact-match in-document entities are less prone to be incorrectly used in generated summaries than OOD entities. From this assumption, a summary-level hallucination risk, called NEHR, can be computed as the percentage of entities in that summary not in the source document. For a document d and a summary s we have:

$$\mathit{NEHR}(d,s) = (1 - \frac{|\mathit{entities} \in d \wedge s|}{|\mathit{entities} \in s|}) \times 100 \quad (1)$$

For detecting named entities and quantities in d and s we rely on an automatic NER system as in (Chen et al., 2021). However, we have no direct way to check if an entity considered as *risky* (*i.e.* not in the source document) is correct or not. Therefore, we rely on human annotations to verify that this criterion is effectively correlated with hallucination errors in generated summaries.

Assessing the relevance of the NEHR criterion The following steps are involved in the empirical study designed to test the relevance of our risk criterion:

1. Select a document/summary corpus C, train several summarization generation models on the training instances of C; generate a set S_d of alternative summaries for all documents d belonging to the test partition of C with the different models and compute the ROUGE and NEHR scores for all summaries $s \in S_d$.
2. For each document $d \in C$, select the maximum ROUGE hypothesis $\hat{s_d}$:

$$\hat{s_d} = \operatorname*{argmax}_{s \in S_d} ROUGE(s, s_{\mathit{ref}})$$

3. Run a NER on each summary $e(\hat{s_d})$.
4. Manually inspect all entities e detected in summaries $\hat{s_d}$ to classify them according to 2 dimensions: inside/outside the source document; correct/incorrect use of e in $\hat{s_d}$.

This study can tell us whether our risk criterion is indeed correlated with the factuality of the summary generated, and whether incorrectly used entities are more frequent outside the document than inside. The motivation for using the max ROUGE summary $\hat{s_d}$ for each document d is to abstract from a given system by using an upper bound of what current state-of-the-art systems can produced. The following section describes the empirical study we did on the CNN/DM corpus.

3 Empirical study on the NEHR criterion

Dataset and model We used the test set of the CNN/DM (Hermann et al., 2015), a popular abstractive dataset for the automatic text summary task that comes from news articles from the CNN and the Daily Mail websites; and BART (Lewis et al., 2020), a transformer encoder-encoder based model that can be fine-tuned to perform automatic text summarization. BART's pre-training consists in applying an arbitrary noising function to corrupt the text deliberately, and training a model to reconstruct the original text from this corrupted version. We initialised the model with pretrained weights from the hugging face library (Wolf et al., 2020) resulting from the fine-tuning of BART-large on CNN/DM[2].

Sampling a diverse population of summaries We generated multiple summaries using 4 different sampling methods to select the next token from a language model: *beam search* that keeps the *num_beams* highest probability paths at each step; *Temperature Sampling* that consists in re-scaling *logits* before applying the softmax; *Top-K Sampling* (Fan et al., 2018) that only keeps the K most likely next words and redistributes the probability among those K words and *Top-P Sampling* (Holtzman et al., 2019) that consists in, given a probability p, taking the smallest possible set of next words whose cumulative probability exceeds a given probability mass and redistributes the probability among them. By using several sampling parameters, we are able to generate a set of 77 summaries for each document to summarize. Refer to appendix A for sampling parameters.

Named entity extraction In this study we use FLERT (Schweter and Akbik, 2020)[3] to extract Named Entities from documents and summaries. FLERT is a NER system that yields 90.93% of F1-score on OntoNotes, a large NER annotated dataset. By extracting NEs on the source documents and on the set of summaries generated by our sampling method on each source document, we are able to estimate the NEHR value for each summary.

Our first evaluation consists of computing the ROUGE metrics on the summary sets generated by our sampling method in order to check that each set contain a large diversity of summaries corresponding to a large range of ROUGE values. We computed ROUGE and NEHR for different summaries produced by the sampling strategy (Table 1). The variations of ROUGE show how the summaries generated cover a large space. Large variations of ROUGE do not translate into large variations of NEHR values for the summaries corresponding to the min and max values of the ROUGE criterion, highlighting the fact that ROUGE might not be a good indicator of summary faithfulness.

summary set	ROUGE (R-1/R-2/R-L)	NEHR
ROUGE max	57.45 / 32.59 /41.63	4.6
ROUGE min	30.04 / 09.33 /19.47	6.0

Table 1: Maximal and minimal values of ROUGE on the summary set produced by our sampling method on CNN/DM. The NEHR value for the min/max summaries is also reported.

	in-doc		out-doc	
Entity dist (%).	79.7		20.3	
Type	*exact.*	*var.*	*inf.*	*hall.*
Type Dist. (%)	62.8	37.2	28.8	71.2
% correct	90	90	88	-

Table 2: % of correctly used entities for each subset of in-document and out of document entities. Dist. is the percentage of entities belonging to each type for entities inside (*exact match, variation*) and outside (*inferable, hallucination*) the document.

Are Out-Of-Document entities mostly hallucinations? We manually analyzed the generated summaries to identify if our assumption that OOD entities were likely to be hallucinations was correct. We randomly selected 50 generated summaries from the test set of CNN/DM with the following constraint: each generated summary must contain at least one exact-match in-document entities and at least one that had no match (either variation of in-document entities or OOD entities). We used these constraints in order to oversample in-document and OOD entities. By using the *exact-match* constraint we were able to select automatically the in-document summaries. In each summary, we manually inspected the same number of entities with exact matches to entities in the source document as those with no matches, resulting in 145 entities with exact matches and 145 with no matches. Three annotators were asked to label each entity as correctly or incorrectly used w.r.t. the following definition: *the entity is used in the correct context according to the document*. The entities inside and outside the document were divided into two types: *exact match* or *variation* for in-document entities and *inferable* or *hallucination* for OOD entities. While *exact match* refers to entities that match exactly those in the source document, *variation* refers

[2] https://huggingface.co/facebook/bart-large-cnn

[3] https://huggingface.co/flair/ner-english-ontonotes-large

to entities in the document that are written with misspellings or using alternative forms, such as a date written differently in the document and in the summary. *Inferable* refers to entities outside the document whose correct use and veracity can be inferred as presented in section 2. We collect the annotations and report the results based on a majority vote among the annotators. The inter-annotators agreement Cohen kappa (Cohen, 1960) of at least 0.63 was obtained for each pair (refer to appendix B for kappa). According to the annotation obtained, most of the annotated entities belong to the set of in-document entities (80%); only 20% belong to OOD entities. In table 2, we report the % of correctly used entities for each subset of in-document and OOD entities. For in-document entities, 90% are labelled as "*correct*", and there are no differences between exact-match and alternative forms of entities. When dealing with OOD entities, 71% of them were considered as hallucinations by our annotators. This confirms our assumption that in-document entities are a good source of information for computing NEHR. It is interesting to see that the set of 29% inferable entities have almost the same correctness (88% *v.s.* 90%) than in-document entities. So the drop in correctness between in-document and out-of-document entities exclusively comes from hallucinations, which represent about 71% of the out-of-document entities. Therefore, by explicitly minimizing the risk of having out-of-document entities, we reduce the risk of hallucinations and this can lead to an increase in summary faithfulness. In the next section we apply the criterion to select summaries at inference time and check the impact on summary quality of explicitly minimizing this criterion.

4 Experiments on summary selection

We evaluate the impact of using our NEHR criterion for selecting a summary at inference time among a possible set of summaries and compare the results obtained in terms of ROUGE and human evaluation with three baseline selection methods: summary with the highest score given by the model among the summaries generated; the 1-best with $beam = 4$ and the summary with the best entailment score compare to the source document as proposed in (Maynez et al., 2020). We propose a criterion based on both NEHR and model scores: First, the population of possible summaries is reduced to those with the lowest NEHR value, then the summary with the highest model score is selected from that subgroup. Let H be the set of summaries sampled from the model, V set of summaries with minimum risk, $P(\cdot|model)$ the probability given by the model to a summary, and $\hat{s}$ the final system output:

$$V = \left\{ x \in H | risk(x) = \min_{s' \in H} NEHR(s') \right\} \tag{2}$$

$$\hat{s} = \operatorname*{argmax}_{s \in V} P(s|model) \tag{3}$$

In section 3, we saw that 37.2% of the entities in the document were variations of other entities belonging to the source document. Thus, using heuristics described in appendix C, we identified entities that were possible variations of in-document entities to ensure that the OOD entities were not mistakenly considered to be alternative forms of in-document entities by our automatic system .

Automatic evaluation We report into table 3, the ROUGE (Lin, 2004), NEHR (see 1) and the percentage of summary with at least one OOD entity computed on CNN/DM. We also report the results of the same experiment made on XSum (Narayan et al., 2018), a more abstractive text summarization dataset than CNN/DM where the reference summaries contain a large number of OOD ngrams: 36% new 1-grams and 83% new 2-grams in XSum whereas CNN/DM has 17% and 54% respectively (Narayan et al., 2018).
The table shows that if BART-Large with beam=4 yields the best automatic evaluation scores, it results in a higher value of NEHR for both dataset. Our approach minimizes that number while maintaining almost equivalent ROUGE for CNN/DM. For XSum, there is a big loss in term of ROUGE. But our proposed method reduce the NEHR as well as the percentage of summary with at least one entity that could be an hallucination. This may be due to the fact that XSum's references summaries are not faithful to the document. In order to check if our selection criterion did not negatively impact the subjective quality of the chosen summary, we performed a manual evaluation on a subsample of the XSum test set.

Human evaluation We evaluated 10 XSum test examples selected randomly. This evaluation consists in annotating the faithfulness of the different entities with respect to the document. Two annotators were asked to label each entity as true if it was

		ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	NEHR ↓	%HallSum ↓
CNN/DM	Beam 4	**43.74**	**20.84**	**30.44**	0.5	3.86
	Best Proba	41.99	18.96	28.01	2.6	20.57
	Entailment	43.61	19.69	29.26	1.62	12.92
	Min NEHR + var (our)	42.19	19.12	28.24	**0.003**	**0.035**
XSum	Beam 4	**45.32**	**22.20**	**37.10**	27.67	52.48
	Best Proba	40.26	16.79	31.29	31.05	61.24
	Entailment	40.92	17.14	31.96	27.08	54.98
	Min NEHR + var (our)	40.16	16.54	31.31	**6.92**	**21.49**

Table 3: Evaluation on CNNDM and XSum. Best Proba – the summary with the best score among all generated summaries given by the model. Min NEHR – our proposed summary selection method after using variation heuristics. %HallSum – the percentage of summary with at least one entity out of the document. ↑ higher is better, ↓ lower is better.

	Annotator. 1	Annotator. 2	Adjudicated
Beam 4	49.33	61.16	48.67
our	59.76	61.07	61.67

Table 4: Average percentage of well-used entities for each system annotated. In adjudicated results, the annotators had to agree on a label for each entity.

used in the right context with respect to the document. An inter-annotator agreement kappa of about 0.38 was obtained. The disagreement among annotators could be attributed to the challenges they faced while annotating entities in sports-related news items where a lot of external knowledge was required to check the correctness of an OOD entity. We held a post-annotation adjudication phase to get annotators to agree on a label for each entity. We report, in Table 4, the percentage of correct entities for each annotator and the adjudication. Looking at the adjudication, 61% of the entities of our method has been tag as correct while 48% for the beam one. That means our method seems to increase the number of correctly used entities by reducing the hallucination risk.

5 Conclusion

We propose in this study a new Named Entity Hallucination Risk criterion for summary selection. Using human evaluation on oracle summaries generated by Bart on CNN/DM, we showed that in-document entities are mostly correct, but this performance drops when considering entities outside documents because of *hallucinations*. We observed empirically that our proposed summary selection method did not significantly impact the ROUGE score for CNN/DM while significantly reducing the hallucination risk. On the highly abstractive dataset XSum, our method was able to drop dramatically the hallucination risk but with a significant drop in ROUGE. Human evaluation of the generated summaries selected from XSum using NEHR showed that the occurring entities were more often correct with respect to those obtained without our selection criteria.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011012525R2 made by GENCI.

References

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. *ArXiv*, abs/1711.04434.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

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.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. 55(12).

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

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

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *ArXiv*, abs/2203.05227.

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.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

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.

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. *arXiv:1910.10683 [cs, stat]*. ArXiv: 1910.10683.

Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition.

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. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

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*.