NonFactS: NonFactual Summary Generation for Factuality Evaluation in Document Summarization

Amir Soleimani
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
a.soleimani.b@gmail.com

Christof Monz
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
c.monz@uva.nl

Marcel Worring
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
m.worring@uva.nl

Abstract

Pre-trained abstractive summarization models can generate fluent summaries and achieve high ROUGE scores. Previous research has found that these models often generate summaries that are inconsistent with their context document and contain nonfactual information. To evaluate factuality in document summarization, a document-level Natural Language Inference (NLI) classifier can be used. However, training such a classifier requires large-scale high-quality factual and nonfactual samples. To that end, we introduce NonFactS, a data generation model to synthesize nonfactual summaries given a context document and a human-annotated (reference) factual summary. Compared to previous methods, our nonfactual samples are more abstractive and more similar to their corresponding factual samples, resulting in state-of-the-art performance on two factuality evaluation benchmarks, FALSESUM and SUMMAC. Our experiments demonstrate that even without human-annotated summaries, NonFactS can use random sentences to generate nonfactual summaries and a classifier trained on these samples generalizes to out-of-domain documents.[1]

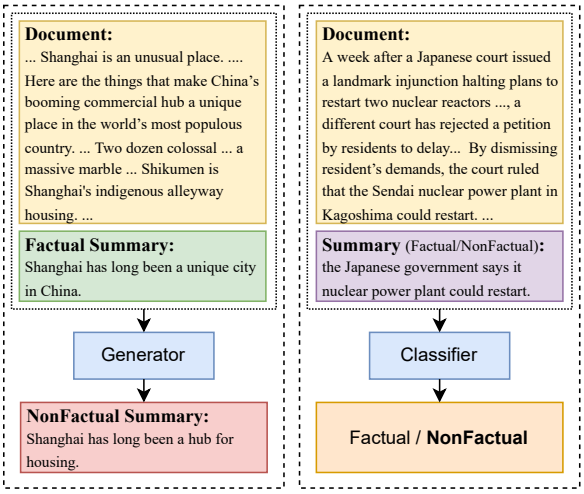

Figure 1: Overview of the proposed pipeline. Left: the NonFactS generator model is trained to generate a nonfactual summary given a reference factual summary and its corresponding context document. Right: Reference factual summaries and the generated nonfactual summaries are used to train a binary classifier to evaluate factuality in document summarization.

1 Introduction

Over the last few years, there have been remarkable improvements in document summarization due to advances in pre-trained language models such as BART and PEGASUS (Lewis et al., 2020; Zhang et al., 2020a). However, these improvements are mainly measured with ROUGE scores, which assess the quality of a summary using n-gram overlap with references. Recent studies show that state-of-the-art models generate up to about 30% nonfactual summaries (Cao et al., 2018; Kryściński et al., 2019; Pagnoni et al., 2021), i.e., summaries that are not entailed by or factually inconsistent with their source document. This demands an automatic evaluation metric for factuality in document summarization.

Factuality evaluation in document summarization is a notoriously difficult task which is closely related to the Natural Language Inference (NLI) task. There have been different attempts to address this problem by revisiting NLI models (Utama et al., 2022; Laban et al., 2021). However, existing NLI datasets such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) do not fully encompass factual inconsistencies within the summarization task. Moreover, NLI datasets cover sentence-level entailment while premises in the summarization task are multi-sentence documents (Utama et al., 2022). On the other hand, NLI approaches need aggregation and, consequently, further in-domain data for training or determining a decision threshold (Laban et al., 2021). In addition,

[1]Codes and Models: github.com/asoleimanib/NonFactS

Findings of the Association for Computational Linguistics: ACL 2023, pages 6405–6419
July 9-14, 2023

collecting human-annotated nonfactual summaries or document-level entailment samples is extremely expensive. Therefore, training a document-level entailment classifier on ground-truth samples is not straightforward because of the lack of data.

A solution to overcome the lack of proper training data is to generate synthetic nonfactual summaries. There have been early attempts to do so using heuristics transformations, e.g., negation, entity swap, and noise injection (Kryscinski et al., 2020), that cover a limited range of possible factual inconsistencies. Recently, FALSESUM (Utama et al., 2022) leveraged a controllable text generation model to replace entity pairs (predicate, argument) in human-annotated reference summaries with new entity pairs. However, it requires extensive pre-processing, impacting the quality of generated samples and results in limited inconsistency variations. Therefore, we extend this line of research to introduce NonFactS, a data generation model to generate nonfactual summaries given a source document and a reference or random summary. We then train a binary classifier on these generated samples to evaluate factuality in document summarization. Figure 1 shows our proposed pipeline, the NonFactS generator and classifier.

NonFactS is trained to complete a truncated reference summary using inputs consisting of only the source document, the truncated reference summary, and a set of random words as $Seeds$. The $Seeds$ are sampled from the document and from the removed part of the summary. In order to generate a nonfactual summary, the $Seeds$ during the inference phase contain random words from the document only. All the words appearing in the reference summary are masked in the document. Figure 2 provides a detailed overview of our generator during training and inference.

The contributions of this work are the following:

First, we introduce a new model to generate nonfactual summaries using a source document and a factual reference summary. Nonfactual summaries are document-level and generated without language-dependent and error-prone pre-processing steps such as entity extraction and lemmatization (see Figure 3).

Second, our method significantly outperforms the state-of-the-art methods on the FALSESUM (Utama et al., 2022) and SUMMAC (Laban et al., 2021) benchmarks.

Third, We demonstrate that our method can still achieve high performance when human-annotated reference summaries are unavailable by using only random sentences from source documents as a substitute.

Fourth, we conduct overlap, novel n-gram, and hypothesis-only analyses to compare NonFactS and FALSESUM regarding their abstractiveness and naturalness of generated summaries.

2 Related Work

This section reviews existing methods for factuality evaluation and standard benchmarks for this task.

2.1 Models

2.1.1 Entity-Based

Laban et al. (2021) introduce a Named Entity Recognition (NER) based method as a baseline to identify if the generated summary entities (e.g., person, location, organizations) are present in the corresponding source document. The quality of NER output significantly impacts the final performance. Dependency Arc Entailment (DAE) (Goyal and Durrett, 2020) is a more advanced model trained on a set of arcs in the dependency parse of generated outputs to classify the entailment decision for each arc with respect to the corresponding input. This approach is also significantly affected by the quality of the parser.

2.1.2 QAG

The Question Answer Generation (QAG) approach follows question generation, question answering, and answer matching steps. FEQA (Durmus et al., 2020) masks text spans (e.g., noun phrases, entities) in the summary, considers the spans as the gold answers, and then generates questions for the gold answers. From there, a Question Answering (QA) model finds answers to these questions in the source documents. F1 performance against the gold answers is considered a faithfulness score. QuestEval (Scialom et al., 2021) combines both a precision-oriented QAG method, with questions generated from the summary such as FEQA, and a recall-oriented metric, with questions generated from the source document such as SummaQA (Scialom et al., 2019). QAG cannot cover all types of factual inconsistency because it significantly depends on entities, and generated questions are mostly factoid.

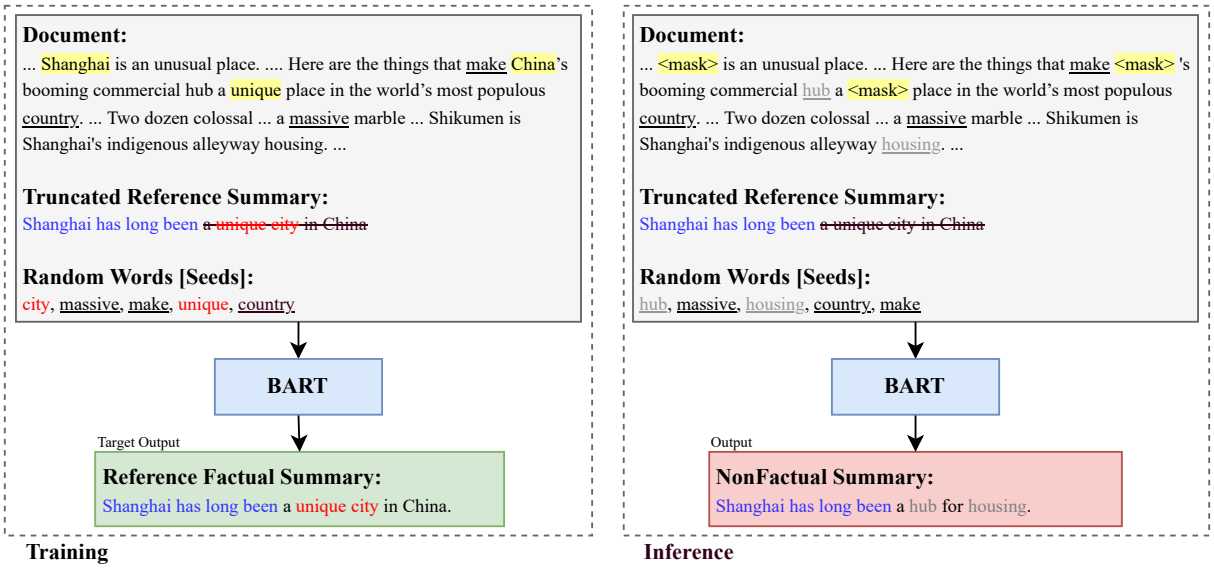

Figure 2: Overview of NonFactS at respectively the training and inference phase. Training: input contains a context document, its truncated reference summary (shown in blue), and random words consisting of words from the document (shown in underline) and from the removed part of the summary (shown in red). The BART model is trained using the reference summaries as targets. Inference: The input structure is the same as the training input but random words are only chosen from the document (shown in underline). In addition the words in the document, which appear in the reference summary, are masked (shown in highlight).

2.1.3 NLI

The NLI task is closely related to factuality evaluation in document summarization. However, premises and hypotheses in the existing NLI datasets such as SNLI and MNLI are sentences while factuality evaluation in document summarization assumes document-sentence pairs. Falke et al. (2019) test five NLI models and compares summaries against all sentences in their corresponding source document and assumes it is sufficient for a summary to be entailed by one source sentence. Laban et al. (2021) introduce a learnable aggregation method and show that their approach outperforms the sentence-level entailment. In general, hypotheses are required to be investigated based on multi-sentence and inter-sentence premises to be classified as entailment, contradiction, or neutral. Furthermore, while mean and max are non-parameter aggregators, learnable methods require additional training data and an in-domain validation set to choose a decision threshold. Document-level entailment pairs solve such challenges.

In order to generate document-level NLI samples, Kryscinski et al. (2020) propose a series of heuristics and rule-based transformations to the sentences of source documents. They introduce a factual consistency checking model (FactCC) that is trained on source documents and the generated sentences pairs. The transformations include paraphrasing to yield semantically-equivalent sentences and negation, pronoun swap, entity swap, number swap, and noise injection to yield semantically-variant sentences. The rule-based nature of the FactCC dataset results in low diversity of factuality errors, and it poorly aligns with actual errors made by summarization models (Goyal and Durrett, 2021).

FALSESUM (Utama et al., 2022) is a data generation pipeline to perturb human-annotated reference summaries. It replaces predicate-argument entities in reference summaries with entities from their corresponding documents. While FALSESUM automatically generates nonfactual summaries, it requires a series of input preprocessing steps (see Figure 3), including entity extraction, span corruption, and lemmatization which are error-prone and language-dependent.

Very recently and concurrently, there have been additional attempts for faithful summarization by automatically generating a synthetic dataset of positive and negative references by corrupting supported reference sentences (Adams et al., 2022) and factual consistency checking by generating factually inconsistent summaries using source texts and reference summaries with key information masked (Lee et al., 2022).

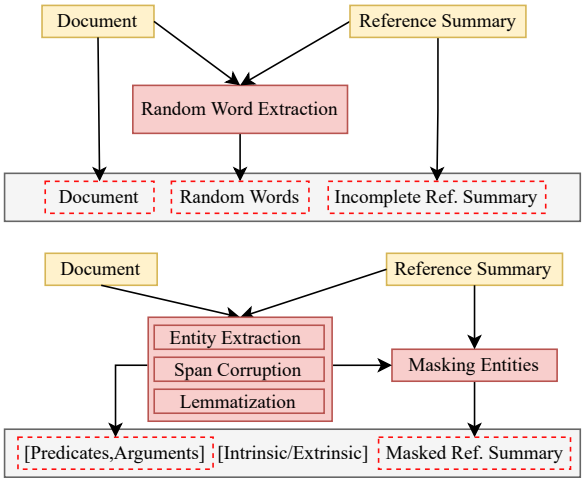

Figure 3: NonFactS and FALSESUM input structures. NonFactS requires only one simple word extraction as pre-processing while entity extraction, span corruption, and lemmatization are needed for FALSESUM.

2.2 Benchmarks

2.2.1 FALSESUM

The FALSESUM benchmark standardizes four manually-annotated datasets: FactCC (Kryscinski et al., 2020), Ranksum (Falke et al., 2019), Summeval (Fabbri et al., 2021), and QAGS (Wang et al., 2020). The dataset labels are imbalanced. Therefore, the performance on the datasets is measured using balanced accuracy (i.e., average recall of two classes) except for Ranksum that uses Precision@1.

2.2.2 SUMMAC

The SUMMAC benchmark comprises the six largest datasets standardized for factuality evaluation: CGS (Falke et al., 2019), XSF (Maynez et al., 2020), Polytope (Huang et al., 2020), FactCC (Kryscinski et al., 2020), SummEval (Fabbri et al., 2021), and FRANK (Pagnoni et al., 2021). SUMMAC also uses balanced accuracy as primary evaluation metric.

3 NonFactS Method

In order to train a classifier to evaluate the factuality of summaries, we need a large set of factual and nonfactual summaries. Reference summaries in large summarization datasets such as CNN (Hermann et al., 2015) and XSUM (Narayan et al., 2018) can be used as factual summaries but the problem is the lack of nonfactual summaries. NonFactS takes a set of source documents D and their corresponding reference factual summaries S^+ and aims to generate a set of nonfactual summaries S^-. The final goal is to train a classifier on pairs of factual and generated nonfactual summaries and their corresponding source documents. S^- should be similar to actual summarizers output and be indistinguishable from S^+ using surface features.

NonFactS is a text generator model taking as an input I, concatenation of D, a truncated factual summary $S^+_{truncated}$, and a list of random words $Seeds$. For training NonFactS, we set $Seeds = \{W_S, W_D\}$, that means random words consist of n random words W_S from $S^+_{removed} = S^+ - S^+_{truncated}$, and m words W_D from D (see Figure 2). It is then trained to generate S^+. In other words, NonFactS is trained to select true words from $Seeds$ to generate a sentence (summary) given the truncated version of that sentence and its corresponding context document. The input format is the following:

$$I = D < /s > S^+_{truncated} < /s > Seeds$$

where $< /s >$ is the separator token. To force the model to generate nonfactual sentences (S^-) at inference time, $Seeds$ are only selected from D ($Seeds = \{W_D\}$), and all the words appearing in S^+ are also masked in D.

The reason to include $S^+_{truncated}$ in the input is to make S^- more indistinguishable from S^+. We set the $S^+_{truncated}$ length to half of the S^+ length that could be the first or last half of the full sentence. In addition, our initial experiments showed if $Seeds$ only contains true words it might result in low quality S^- as the model has to complete $S^+_{truncated}$ using all words, which can be completely irrelevant words to $S^+_{truncated}$. Therefore, we include more words than needed in $Seeds$ to force the model to select more suitable words. Note, $Seeds$ contains only half of $S^+_{removed}$ words to encourage the model to use the context information in D. Words are shuffled and the set does not contain stop words.

We use BART-base (Lewis et al., 2020) as our generator model and use the CNN summarization dataset as our training dataset. The training set has more than 287k samples from which we randomly choose 50k samples for the inference phase. We split summaries into sentences which results in about 900k training pairs (document, sentence). We use a batch size of 40 samples and a learning rate of $3e10^{-5}$, and train the model for one epoch on 2 NVIDIA TITAN X Pascal GPUs (12GB memory)

Document: Thousands on Saturday fled the area in southwestern Ivory Coast where attacks left seven U.N. peacekeepers and eight civilians dead, according to a U.N. official. ... Humanitarian organizations reported Saturday they were expecting about 4,000 people in Tai, said Remi Dourlot, a spokesman for the U.N. Office for the Coordination of Humanitarian Affairs. ... U.N. Operation in Cote d'Ivoire and Ivory Coast troops have increased their presence in the area, Dourlot said Saturday. ... **Reference Factual Summary:** Humanitarian groups expect 4,000 refugees in one camp, a U.N. official says. **Half Summary + Seeds:** xhumanitarian groups expect 4,000 refugees in </s> understood + accountable + Ivoire + attacks + included + west + expecting + seven + volunteers + armed + occurred + Dourlot + Cote + reasons **Generated NonFactual Summary:** Humanitarian groups expect 4,000 refugees in Cote d'Ivoire, U.N. spokesman says.
Document: For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals – and they come from all over the world. ... That doesn't mean Francis is the first pontiff to appoint cardinals from the developing world, though. **Reference Factual Summary:** The 15 new cardinals will be installed on February 14. **Half Summary + Seeds:** be installed on February 14. </s> canonized + reach + Kean + number + like + pontiff **Generated NonFactual Summary:** The new pontiff will be installed on February 14.
Document: Rebels in Tripoli furiously hunting for signs of longtime Libyan leader Moammar Gadhafi are exploring a network of tunnels and bunkers built beneath his massive compound. CNN's Sara Sidner got a peek at the passageways Friday. She dubbed it "Gadhafi's inner sanctum." ... **Reference Factual Summary:** CNN's Sara Sidner sees another world in a tunnel below Tripoli. **Half Summary + Seeds:** world in a tunnel below Tripoli. </s> extend + walked + underground + shelf + occurred + thought + apparently + passages + air + recently **Generated NonFactual Summary:** Rebels are exploring underground passages around the world in a tunnel below Tripoli.
Document: Criminals who file fraudulent tax returns by stealing people's identities could rake in an estimated 26 billion... But in testimony before Congress last year, National Taxpayer Advocate Nina Olson said those filters "inevitably block large numbers of proper refund claims" since there "is no easy way to distinguish proper claims from improper ones." In testimony prepared for Tuesday's hearing, Deputy IRS Commissioner Steven Miller said the agency cannot stop all identity theft. ... **Reference Factual Summary:** The Treasury's estimate is the first detailed analysis of the ongoing problem. **Half Summary + Seeds:** the Treasury's estimate is the first </s> detects + numbers + billion + 6 + cars + Security + agency + recently + Congress + 5 **Generated NonFactual Summary:** The Treasury's estimate is the first to be presented to Congress by the agency.

Table 1: Examples of NonFactual summaries generated by the NonFactS generator. Documents are truncated for visibility. Note, Reference Factual Summary is not an input for the model and presented for comparison.

	FALSESUM Benchmark Datasets				
Dataset	**FactCC**	**Ranksum**	**QAGS**	**SummEval**	**Overall**
MNLI	57.9	51.4	52.7	48.8	51.4
ANLI*	53.9	55.8	53.5	49.6	53.2
DocNLI*	58.1	53.6	57.1	52.6	55.4
FactCC*	73.9	67.3	73.5	60.0	69.0
FALSESUM*	83.5	72.9	**75.1**	65.2	74.2
NonFactS 100k	84.2	77.6	70.7	71.2	75.9
NonFactS* 100k	**86.2**	**77.8**	72.5	**72.3**	**77.2**

Table 2: FALSESUM benchmark (Utama et al., 2022). *: training dataset is augmented with the MNLI dataset.

for about one day. Table 1 shows four nonfactual summaries generated by the NonFactS generator.

To evaluate the factuality of generated summaries, we choose ROBERTa (Liu et al., 2020) and ALBERT (Lan et al., 2020) as our default classification models and fine-tune the models on a balanced dataset consisting of generated nonfactual summaries, reference factual summaries, and context documents ($S = \{S^+, S^-\}, D$).

4 Experiments

4.1 Benchmark Results

We evaluate NonFactS on two factuality evaluation benchmarks, FALSESUM and SUMMAC. Performance is measured using Balanced Accuracy (BA):

$$BA = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$$

where TP, FN, TN, and FP stand for true positive, false negative, true negative, and false positive, respectively. The majority performance for BA is 50.

Model	SUMMAC Benchmark Datasets						Overall
	CGS	XSF	Polytope	FactCC	SummEval	FRANK	
NER-Overlap	53.0	63.3	52.0	55.0	56.8	60.9	56.8
MNLI-doc	57.6	57.5	61.0	61.3	66.6	63.6	61.3
FactCC-CLS	63.1	57.6	61.0	75.9	60.1	59.4	62.8
DAE	63.4	50.8	62.8	75.9	70.3	61.7	64.2
FEQA	61.0	56.0	57.8	53.6	53.8	69.9	58.7
QuestEval	62.6	62.1	**70.3**	66.6	72.5	82.1	69.4
$\text{SUMMAC}^{\dagger}_{ZS}$	70.4	58.4	62.5	83.8	78.7	79.0	72.1
$\text{SUMMAC}^{\dagger}_{conv}$	64.7	**66.4**	62.7	89.5	81.7	81.6	74.4
$\text{FALSESUM}^{\dagger}$	74.7	51.1	63.7	87.7	86.8	80.0	74.0
$\text{NonFactS}^{\dagger}$	81.6	53.2	60.8	89.3	87.4	80.1	75.4
$\text{NonFactS}^{\dagger\dagger}$	**81.7**	54.0	61.2	**90.6**	**89.0**	**84.3**	**76.8**

Table 3: SUMMAC benchmark (Gliwa et al., 2019). †: ALBERT-xlarge and ††: ALBERT-xxlarge.

Dataset	R-1	R-2	R-3	R-4	R-L	BERTScore (F1)
NonFactS	58.6	49.2	43.0	36.3	58.2	56.6
FALSESUM	54.2	42.0	34.3	27.9	53.5	55.0

Table 4: ROUGE scores between positive and negative samples in NonFactS and FALSESUM. Negative samples in NonFactS are more similar to their positive pairs in terms of ROUGE scores and BERTScore.

Model	FALSESUM	NonFactS
Majority voting	50.00	50.00
RoBERTa-base	69.31	68.53
RoBERTa-large	73.54	72.13

Table 5: Hypothesis-only model performance. Models are trained on 80% of the training set and evaluated on the remaining 20% samples. Lower is better.

Train / Test	NonFactS	FALSESUM
NonFactS	-	80.73
FALSESUM	78.39	-

Table 6: Comparing NonFactS and FALSESUM on identifying synthetic samples.

Table 2 reports NonFactS's performance on the FALSESUM benchmark. For this benchmark, ROBERTa-base is fine-tuned on 100k factual/nonfactual samples augmented with MNLI. NonFactS outperforms overall performance on all datasets except QAGS. It also reports NonFactS without augmentation data and shows that it outperforms FALSESUM. QAGS categorizes non-grammatical sentences as non-consistent (nonfactual) (Wang et al., 2020). We also manually investigated QAGS and found numerous non-grammatical, but factually correct, sentences labelled as nonfactual samples. We suspect that such a phenomenon and the fact that we generate grammatically correct sentences only might be the reason for our seemingly lower performance on QAGS.

Table 3 compares different models' performance on the SUMMAC benchmark. The experimental setup in this benchmark does not limit the number of training samples and the size and type of the classification model. We fine-tune ALBERT (Lan et al., 2020) on our 200K balanced datasets. The SUMMAC model uses ALBERT xlarge and larger datasets (MNLI and VitaminC (Schuster et al., 2021)). NonFactS outperforms the overall balanced accuracy performance. It is also considerably better on the CGS and SumEval datasets but performs poorly on XSF. We manually investigated XSF and suspect that the poor performance of our model and other models might be because of the high frequency of non-grammatical, noisy, and nonsense sentences labelled as nonfactual (e.g., 'barron and his wife barron have moved from the white house to the white house'). It is also understandable from the NER-Overlap model, which is the second-best model on XSF compared to the much more advanced models. In contrast to other datasets, XSF was mainly collected from the XSUM dataset. While this domain shift can be a reason for the low performance, this is not the case for our model. We experimented with NonFactS trained on our synthetic dataset based on XSUM and did not see a significant improvement.

4.2 Fine-grained Analysis

In order to have high quality nonfactual samples for training a binary classifier, nonfactual samples must not be identified by surface features. Table 4 compares NonFactS and FALSESUM regarding the similarity of factual and generated nonfactual samples. NonFactS's nonfactual samples are much

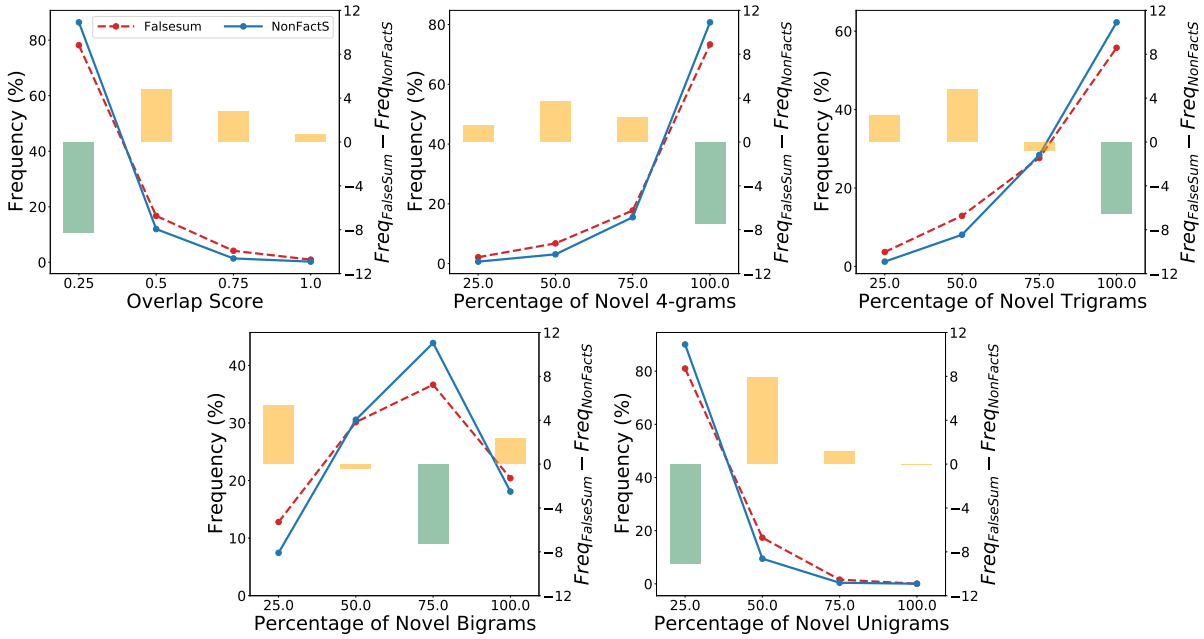

Figure 4: Comparing NonFactS and FALSESUM synthetic samples. Average percentage of novel n-grams [FALSESUM,NonFactS]: 4-grams mean=[83,87], trigrams=[73,77], bigrams=[52,54], unigrams=[13,10]. Bar charts show $Frequency_{FALSESUM} - Frequency_{NonFactS}$. Green:$Frequency_{FALSESUM} < Frequency_{NonFactS}$.

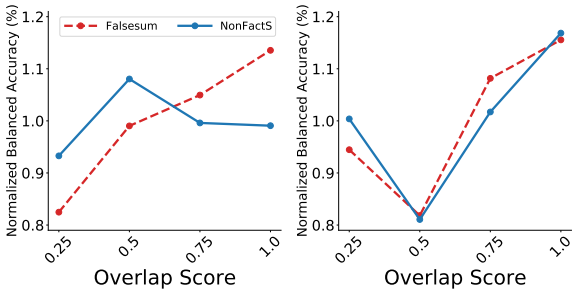

Figure 5: Comparing NonFactS and FALSESUM performance on the FALSESUM (left) and SUMMAC (right) benchmarks regarding to different levels of overlap between summaries and their documents.

more similar to factual samples in terms of ROUGE scores and BERTScore (Zhang et al., 2020b). In addition, inspired by Gururangan et al. (2018) and Utama et al. (2022), we perform a hypothesis-only experiment. The classifier is trained and evaluated on only summaries without any access to the context documents. The goal is understanding to what extent the factuality of generated summaries can be determined using semantic plausibility and spurious surface features (e.g., grammatical mistakes or fluency errors). Table 5 indicates that NonFactS generated summaries are marginally better than FALSESUM generated summaries in hypothesis-only factuality evaluation. We also manually investigated 100 randomly sampled generated nonfactual summaries and found that 85% of the labels are truly labelled as nonfactual. This is almost the same as FALSESUM reported manual verification (Utama et al., 2022).

We study the ability of the same classifier (ALBERT xlarge) fine-tuned on the NonFactS/FALSESUM datasets to evaluate factuality on FALSESUM/NonFactS. The rest of the variables, such as the number of training samples, are the same as our default. Table 6 indicates that NonFactS yields better performance on FALSESUM.

We investigate the performance of the NonFactS factuality evaluation model based on the level of abstractiveness of summaries. We use different metrics to partition the lexical overlap between the summaries and their context documents. Overlap Score is defined by the multiplication of the density, i.e., the percentage of words in a summary that are present in the context document, and normalized coverage, i.e., the percentage of a summary that is a continuous fragment of the context document (Utama et al., 2022; Grusky et al., 2018). We also use the percentage of novel n-grams in summaries, i.e., the percentage of a summary n-grams that are not present in the context document. Higher values for the overlap score and lower values for percentage of novel n-grams correspond to higher overlap and more extractive summaries.

Figure 4 plots NonFactS and FALSESUM re-

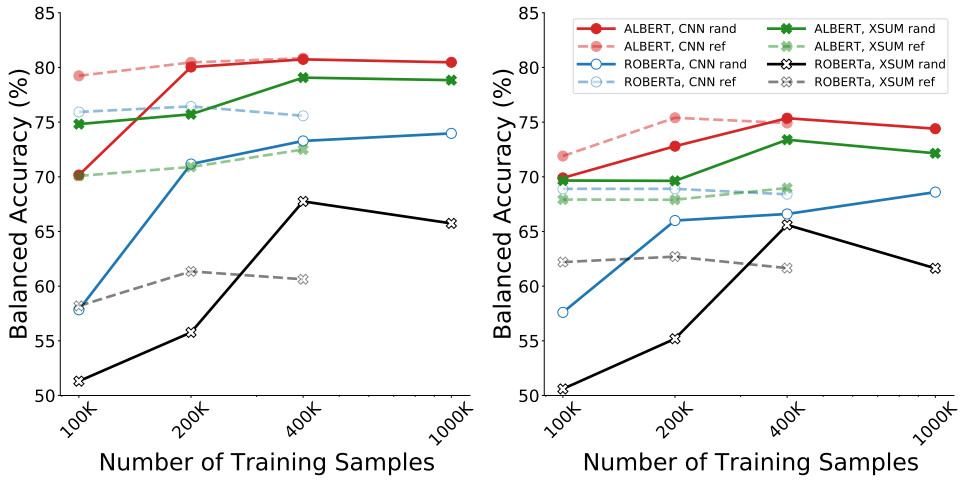

Figure 6: Performance of NonFactS in absence of reference summary on the FALSESUM (left) and SUMMAC (right) benchmarks. Random sentences from documents are used as factual summaries, and nonfactual summaries are generated using random sentences. This experiment is not an extractive summarization approach since random sentences are removed from documents.

garding the overlap score and percentage of novel n-grams. Both generated datasets cover more abstractive than extractive summaries. However, NonFactS contains more abstractive samples. This is evident from the higher frequency of lower overlap scores. NonFactS also has more samples with a higher percentage of novel 4-grams and trigrams, while FALSESUM covers more novel Bigrams and unigrams. To study the effect of summary extractiveness, we evaluate our model on the FALSESUM and SUMMAC benchmarks. Figure 5 indicates the higher performance of NonFactS over FALSESUM on more abstractive summaries (lower overlap scores) on both benchmarks, which is in line with more abstractive samples in NonFactS.

4.3 Zero Reference Analysis

In this section, we consider the case in which there is no access to human-annotated reference summaries (factual summaries) for training a model to generate nonfactual summaries. This is a realistic case, for example, in a real scenario where one has no access to reference summaries in a new domain.

We use randomly selected sentences from context documents as factual reference summaries corresponding to the documents. Next, we train the NonFactS generator with the same procedure explained in Section 3 to generate nonfactual summaries. Note, during the training and inference phase, we remove the randomly selected sentences from the documents to eliminate trivial performance and maintain the abstractive summarization approach. The exact number of documents (230k/50k) are used for training and inference. Documents during inference are sampled more than once to provide more samples (200k,400k,1000k) for training the classifier.

To single out the model and dataset effects, we experiment with both ROBERTa and ALBERT and CNN and XSUM as training and inference datasets. The default case (presence of reference summary) is limited regarding the number of training samples for the classifier (max 400k samples).

Figure 6 compares the performance of the factuality evaluation models in the presence and absence of reference summaries on the FALSESUM and SUMMAC benchmarks (see Appendix for detailed results). In both benchmarks, zero reference models reach or outperform reference models after training on 400k random factual samples and their corresponding nonfactual summaries. This superiority is much more evident in the ALBERT models. In addition, the figure shows that CNN based models performs better on both benchmarks which is to be expected as both benchmarks are consisting of more CNN based datasets. However, we see that the ALBERT models trained on CNN or XSUM random samples relatively converge together. Therefore, the effect of in-domain datasets vanishes as the model trained on more samples.

5 Conclusion

We introduced NonFactS, a data generation model to generate large-scale nonfactual summaries. Non-

FactS only requires context documents and reference summaries as factual summaries. To evaluate factuality in document summarization, we used a binary classifier trained on a balanced dataset of factual and generated nonfactual summaries. Our model outperforms prior works on two standard benchmarks, FALSESUM and SUMMAC.

Compared to previous methods, NonFactS generates nonfactual samples without requiring extensive language-dependent pre-processing steps. Also, our generated samples are more abstractive and more similar to their factual references, and therefore, it is harder to identify the samples based on spurious surface features and semantic plausibility.

Additionally, we demonstrated that NonFactS is capable of generating nonfactual summaries without the need for human-annotated reference summaries by utilizing randomly selected sentences from context documents. Our experiments indicated that a classifier trained on these generated samples achieves comparable performance to a classifier trained on human-annotated samples and their generated nonfactual pairs.

Limitations

NonFactS generates grammatically correct nonfactual summaries. However, in practice, summaries can be non-grammatical, noisy, and nonsensical. This can limit the generalization of our performance in such cases. Additionally, hypothesis-only results show that a considerable number of samples are identified correctly without their context document. The reason can be the memorized knowledge in pre-trained classifiers or surface features and semantic plausibility.

Broader Impact

Our model has no direct environmental impacts, fairness or privacy considerations. However, it is important to note that it must not be used as a fact-checking tool as there is a potential risk that false statements may be labelled as true. Our classifier evaluates the factuality of a summary based on a context document, and if the document is misleading, the summary can be factual based on misleading information. Additionally, NonFactS generates nonfactual summaries, which might have potential risks if misused for generating massive nonfactual summaries (claims). Addressing such risks is an open issue in the field and is not specific to our work.

Acknowledgments

This research was partly supported by Athora Netherlands and the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080.

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *AAAI*, pages 4784–4791.

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.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. 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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *CoRR*, abs/2111.09525.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee, and Kyomin Jung. 2022. Masked summarization to generate factually inconsistent summaries for improved factual consistency checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1019–1030, Seattle, United States. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

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.

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.

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.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. 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 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

Table 7 and 8 show the detailed results for comparing the performance of our model with and without access to human-annotated factual summaries (see Figure 6).

Dataset	FALSESUM Benchmark Datasets				Overall
	FactCC	Ranksum	QAGS	SummEval	
ROBERTa, CNN reference summaries					
100k	84.2	77.6	70.7	71.2	75.9
200k	84.0	77.1	73.9	70.7	76.4
400k	81.1	77.9	73.7	69.7	75.6
ROBERTa, CNN random sentences					
100k	63.8	54.2	50.7	62.7	57.8
200k	79.1	60.8	74.2	70.5	71.2
400k	79.9	68.3	74.5	70.4	73.3
1000k	80.3	70.9	73.8	70.9	74.0
ALBERT, CNN reference summaries					
100k	87.2	79.2	75.0	75.5	79.2
200k	87.8	79.8	76.7	77.5	80.5
400k	86.2	79.5	79.8	78.0	80.9
ALBERT, CNN random sentences					
100k	82.1	78.3	74.7	73.6	77.2
200k	87.2	79.1	79.7	74.3	80.0
400k	88.0	79.2	79.8	76.0	80.7
1000k	87.0	78.9	79.4	76.6	80.5
ROBERTa, XSUM reference summaries					
100k	60.8	51.6	56.6	63.7	58.2
200k	66.0	55.2	60.5	63.6	61.4
400k	63.3	55.5	60.8	62.9	60.6
ROBERTa, XSUM random sentences					
100k	54.3	50.7	48.1	52.1	51.3
200k	61.1	52.4	53.4	56.3	55.8
400k	75.0	56.8	71.0	68.1	67.8
1000k	88.0	71.2	79.2	77.0	78.8
ALBERT, XSUM reference summaries					
100k	76.3	60.2	71.6	72.2	70.1
200k	77.0	61.5	72.8	72.3	70.9
400k	79.7	65.8	72.0	72.6	72.5
ALBERT, XSUM random sentences					
100k	82.5	68.6	74.1	74.1	74.8
200k	83.6	69.5	74.7	75.1	75.7
400k	88.1	71.4	78.6	78.2	79.1
1000k	88.0	71.2	79.2	77.0	78.8

Table 7: Comparing the performance of NonFactS with and without human-annotated reference summaries on FALSESUM. In the absence of human-annotated samples, random sentences from documents are used as factual summaries, and nonfactual summaries are generated using random sentences.

Dataset	SUMMAC Benchmark Datasets						Overall
	CGS	XSF	Polytope	FactCC	SummEval	FRANK	
ROBERTa, CNN reference summaries							
100k	69.2	48.6	59.2	83.0	78.3	74.8	68.9
200k	72.9	49.5	55.6	82.3	79.7	73.1	68.9
400k	69.6	50.7	57.0	88.8	72.4	73.3	68.6
ROBERTa, CNN random sentences							
100k	55.6	51.4	48.8	60.4	58.1	71.4	57.6
200k	63.5	48.8	55.0	78.6	73.6	76.3	66.0
400k	63.9	53.3	52.5	81.8	72.4	75.6	66.6
1000k	69.9	53.4	58.5	78.9	78.2	74.9	68.6
ALBERT, CNN reference summaries							
100k	72.8	52.7	57.1	88.3	83.3	77.9	71.9
200k	81.6	53.2	60.8	89.3	87.4	80.1	75.4
400k	79.9	54.2	60.8	87.0	85.9	81.9	74.9
ALBERT, CNN random sentences							
100k	63.7	50.6	62.4	83.9	74.9	82.7	69.7
200k	74.2	49.1	61.1	88.5	83.3	80.9	72.8
400k	78.0	52.0	61.4	86.3	91.6	82.8	75.4
1000k	77.2	52.4	60.9	85.3	88.6	82.0	74.4
ROBERTa, XSUM reference summaries							
100k	60.9	53.7	55.8	58.4	70.0	74.6	62.2
200k	57.8	54.7	53.4	65.8	72.2	72.3	62.7
400k	59.2	54.0	49.9	61.6	73.2	72.0	61.7
ROBERTa, XSUM random sentences							
100k	50.1	49.0	49.4	53.8	51.1	50.2	50.6
200k	52.5	51.3	48.4	58.6	55.9	64.6	55.2
400k	56.5	59.6	59.9	74.0	68.4	75.2	65.6
1000k	56.0	48.9	57.5	69.3	64.1	74.1	61.6
ALBERT, XSUM reference summaries							
100k	67.5	50.5	57.3	74.6	80.8	81.4	67.9
200k	63.6	50.9	57.8	75.5	78.2	81.4	67.9
400k	68.1	52.1	57.3	77.3	77.3	81.9	69.0
ALBERT, XSUM random sentences							
100k	67.2	51.7	56.4	83.7	78.5	80.5	69.7
200k	68.2	48.4	59.3	83.6	77.2	81.2	69.6
400k	72.1	52.7	60.0	90.5	82.7	82.4	73.4
1000k	69.2	52.4	59.5	87.8	81.6	82.5	72.2

Table 8: Comparing the performance of NonFactS with and without human-annotated reference summaries on SUMMAC. In the absence of human-annotated samples, random sentences from documents are used as factual summaries, and nonfactual summaries are generated using random sentences.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation section just after conclusion

☑ A2. Did you discuss any potential risks of your work?
Broader Impact section just after Limitation section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract. Introduction (section 1)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The datasets we used are the two well-known datasets in the field and are publicly available and free to use for academic and research purposes. When we publish the dataset, we will provide the licence and respect the previous licences.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
CNN and XSUM summarization datasets are public and free to use for academic and research purposes. We completely respect their intended use. When we publish the dataset, we will provide the licence and respect the previous licences. We have considered that our contribution is compatible with the original datasets.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. It is not applicable since the data we use have already been checked by their authors.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. It is completely the same as the datasets we use and therefore we only cite corresponding works.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3. We stick to default models parameters and for our specific parameters, we discussed that.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. We do not need specific packages but when we publish our model we will specify frameworks and other requirements versions.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.