Multi-View Source Ablation for Faithful Summarization

Shuyang Cao[1*] **Liang Ma**[2] **Di Lu**[2]
Robert L. Logan IV[2] **Joel Tetreault**[2] **Alejandro Jaimes**[2]
[1]University of Michigan, Ann Arbor [2]Dataminr Inc.
caoshuy@umich.edu {lma,dlu,rlogan,jtetreault,ajaimes}@dataminr.com

Abstract

In this paper, we present MuFaSSA (Multi-view Faithfulness Scoring via Source Ablation), a metric for *evaluating* faithfulness of abstractive summaries, and for guiding *training* of more faithful summarizers. For evaluation, MuFaSSA employs different strategies (e.g., masking entity mentions) to first remove information from the source document to form *multiple ablated views*. Then, the faithfulness level of each token in a generated summary is measured by the difference between the token generation probabilities when given the original document and the ablated document as inputs to trained summarizers. For training, MuFaSSA uses a novel *word truncation* objective that drops unfaithful tokens located by MuFaSSA in both the decoder input and output. Alignments with human-annotated faithfulness labels on AggreFact show that MuFaSSA is comparable to or better than existing metrics built on classifiers or QA models pre-trained on other tasks. In experiments on summarization with XSum and CNN/DailyMail, models trained with word truncation using MuFaSSA outperform competitive methods according to both automatic faithfulness metrics and human assessments.

1 Introduction

Automatic text summarization systems have made great strides with the use of large pre-trained models, which enable more precise identification of salient content in the document and generation of summaries with human-level fluency (Lewis et al., 2020; Raffel et al., 2020). However, model-generated summaries frequently contain unfaithful information that either contradict the source text or cannot be verified (Kryscinski et al., 2020), creating risks in real-world deployment of automatic text summarization models and motivating the development of models targeting more faithful summaries (Cao et al., 2018).

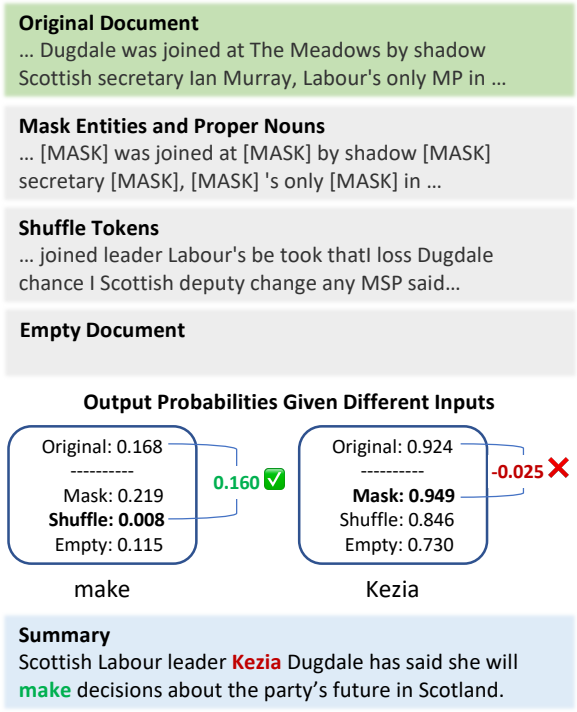

Figure 1: MuFaSSA estimates the faithfulness level of each summary token as the difference between probabilities given by trained summarizers with the original source document and the *ablated document* chosen for the token (e.g., shuffling tokens for verbs). The large difference for "make" indicates it is **faithful** to the document, while the small difference for "Kezia" indicate an **unfaithful** token.

As overlap-based metrics such as ROUGE (Lin, 2004) struggle to reflect the faithfulness level of generated summaries (Falke et al., 2019), a number of model-based faithfulness metrics have been introduced. These metrics leverage external textual entailment (Goyal and Durrett, 2020; Laban et al., 2022) and question answering models (Wang et al., 2020; Scialom et al., 2021) to measure the degree to which claims in the summary align to information in the source text. Yet, there remains substantial room for improvement (Tang et al., 2022). Moreover, despite being relevance or complementary to

[*] Work done during an internship at Dataminr.

Findings of the Association for Computational Linguistics: EACL 2023, pages 2029–2047

each other, for building faithful summarization systems, faithfulness metrics are rarely exploited and researchers mainly resort to more complex training routines (Cao and Wang, 2021) or model architectures (Zhu et al., 2021).

To this end, our work introduces a faithfulness metric that (1) more accurately estimates summary faithfulness levels; and (2) can be easily integrated into training objectives to produce more faithful summarization systems. In our method, which we call **MuFaSSA** (**Mu**lti-view **Fa**ithfulness **S**coring via **S**ource **A**blation), multiple *ablated source documents* are constructed by masking entities, shuffling tokens, and discarding all tokens in the original source document, to remove crucial information for the generation of different content in the summary, as shown at the top of Figure 1. Since the ablated sources do not include sufficient information for generating the corresponding summary, the differences between token output probabilities given by the original input and each ablated input should be high for the faithful content and low for the unfaithful one (e.g., "make" and "Kezia" in Figure 1). We then aggregate the differences to obtain the summary faithfulness score.

Additionally, to train faithful summarization systems, we adapt loss truncation (Kang and Hashimoto, 2020) and nullify losses on summary tokens that are deemed less faithful by MuFaSSA during training. Compared to using training losses for detecting unfaithful content in the original loss truncation, MuFaSSA provides a more accurate estimation of token faithfulness in training samples, and more faithful summaries can therefore be produced while maintaining informativeness. We further design **word truncation**, to drop the generation dependency on less faithful words in the auto-regressive decoder by completely removing them from the decoder input.[1]

Two sets of experiments are conducted to show the effectiveness of MuFaSSA at evaluating and training faithful summarizers. First, we compare with existing faithfulness metrics on AggreFact (Tang et al., 2022), a curated benchmark for meta evaluation of faithfulness metrics, where MuFaSSA obtains the best average performance. We then leverage MuFaSSA during model training on XSum (Narayan et al., 2018) and CNN/DailyMail (Hermann et al., 2015). Compared to baselines and recent models built with augmented data or more complex training objectives, MuFaSSA-trained models produce summaries with competitive or better faithfulness while maintaining the coverage of salient document information, according to both automatic faithfulness metrics and human judgments.

In summary, we make the following contributions:

- We propose MuFaSSA, a novel automatic evaluation metric that measures summary faithfulness by the extent to which the generation of summary tokens relies on information in the document.
- In addition, to leverage MuFaSSA during training, we design word truncation, a novel training objective that discards less faithful tokens identified by MuFaSSA from the training samples to induce more faithful summarizers.

2 Related Work

Faithfulness Metrics. Recent analyses have shown that summaries with high ROUGE scores (Lin, 2004) can contain information that is not faithful to the source documents (Falke et al., 2019; Kryscinski et al., 2020). This observation has prompted the development of a number of faithfulness metrics that measure the extent to which summarization models produce unfaithful outputs. Existing faithfulness metrics largely fall under two broad categories: (1) entailment-based metrics, and (2) QA-based metrics. Entailment-based metrics evaluate the faithfulness of summaries by computing entailment levels of the sentences (Laban et al., 2022), dependency arcs (Goyal and Durrett, 2020), or semantic graphs (Ribeiro et al., 2022) of the summaries against the corresponding documents. QA-based metrics use models for question generation and question answering to determine whether questions derived from the summary can be answered using the document (Wang et al., 2020; Durmus et al., 2020) or questions derived from the document can be answered using the summary (Scialom et al., 2019). Results are enhanced using a combination of both approaches (Scialom et al., 2021) and adding a question filtering stage (Fabbri et al., 2022).

In this work we pursue an alternative approach that detects unfaithful outputs by analyzing differences in token probabilities when conditioning on different views of the source document. This

[1] Our code is available at https://shuyangcao.github.io/projects/mufassa/.

approach was first proposed by Xie et al. (2021), whose CoCo metric measures probability differences on pre-specified sets of key terms in the output of models conditioned on partially masked sequences. Our proposed metric, MuFaSSA, builds upon CoCo in two ways. First, we eschew the need for key terms, instead providing an approach for assessing faithfulness of different types of tokens (e.g., entities, relations). This not only makes MuFaSSA easier to use, but, as we will see in Section 4, also results in better performance. Secondly, we introduce a strategy for incorporating these metrics into training, and demonstrate in Section 5.2 that this training scheme produces more faithful summarizers.

Faithful Summarization. In parallel with advancements in faithfulness metrics, researchers have also investigated approaches to train more faithful summarizers. One class of approaches propose to modify model architectures to leverage external knowledge graphs (Zhu et al., 2021) and OpenIE triplets (Cao et al., 2018). Another class of approaches investigates modifications to training data, either by removing unfaithful training examples (Wan and Bansal, 2022) or training models to differentiate between faithful and unfaithful summaries (Liu et al., 2021; Cao and Wang, 2021). In this paper, we study a third class of approaches that modify the model's loss function. Our work builds upon the method of loss truncation (Kang and Hashimoto, 2020), which omits a fraction of low confidence predictions from the loss function during training. We show that loss truncation can better improve faithfulness by using MuFaSSA to determine which predictions to ignore, and that even better results can be obtained using our novel word truncation objective that omits removed tokens from the input (Tables 2 and 3).

3 MuFaSSA: Multi-View Information Ablation

In this section, we first introduce the formulation of faithfulness estimation by MuFaSSA (§3.1) and the construction of inputs with different information ablated (§3.2). We then describe how MuFaSSA can be incorporated into model training through loss truncation and our proposed word truncation (§3.3). We fine-tune BART (Lewis et al., 2020) for all summarization models in this paper.

3.1 Information Ablation

Let T denote the set of tokens in the model vocabulary, and T^* the set of all sequences of tokens in T. Given a summary $y \in T^*$ of document $x \in T^*$, let $\mathcal{I}_{y_i} : T^* \rightarrow T^*$ denote a "view function" that ablates out information from the source document necessary for generating token y_i (i.e., a summarization model conditioned on $\mathcal{I}_{y_i}(x)$ should *not* produce token y_i). We hypothesize that if y_i is not faithful to the source document, then y_i can be generated with $\mathcal{I}_{y_i}(x)$ by a summarization model, i.e., the output probability $p(y_i|y_{<i}, \mathcal{I}_{y_i}(x))$ should remain high. Based on this hypothesis, we propose the faithfulness level of summary token y_i estimated by:

$$m(y_i) = p_\theta(y_i|y_{<i}, x) - p_{\theta'}(y_i|y_{<i}, \mathcal{I}_{y_i}(x)) \quad (1)$$

where a higher $m(y_i)$ suggests a higher faithfulness level, and p_θ and $p_{\theta'}$ denote summarization models parameterized by θ and θ'. We train p_θ and $p_{\theta'}$ on the experimented summarization dataset by maximizing $p_\theta(y_i|y_{<i}, x)$ and $p_{\theta'}(y_i|y_{<i}, \mathcal{I}_{y_i}(x))$ with the cross-entropy objective. To obtain the sample-level faithfulness score, we aggregate the faithfulness estimation over all summary tokens: $\frac{1}{L}\sum_{i=1}^{L} m(y_i)$, where L is the length of the summary. Notably, the token-level faithfulness scores aggregated by MuFaSSA are based on the generation probabilities given the *already generated tokens*. The contextual nature of the token-level faithfulness scores allows MuFaSSA to account for unfaithfulness of phrases and sentences in the generated summary.

3.2 Multi-View Ablation

Careful construction of $\mathcal{I}_{y_i}$ is crucial to accurate faithfulness estimation of y_i. To reduce the computational cost, instead of creating a unique ablated document $\mathcal{I}_{y_i}(x)$ for every y_i, MuFaSSA groups the summary tokens into three different sets—Y_{ent}, Y_{rel}, and Y_{other}—according to their part-of-speech (POS) tags and entity labels,[2] and constructs a single view of the source document I_Y to compute $m(y_i)$ for each token $y_i \in Y$.[3] In the following paragraphs, we describe the construction strategies and their corresponding token sets.

[2] We use SpaCy (Honnibal and Montani, 2017) for named entity recognition and part-of-speech (POS) tagging.

[3] Thus, all of the token-level scores are computed using only 4 forward passes of the model: one for each set, and one for the original source document.

Mask Entities and Proper Nouns. Named entities and proper nouns are important components of facts and events that constitute summaries, so our first set of tokens, Y_{ent} is comprised of all tokens that are part of a proper noun or named entity. Since the faithful production of these tokens in the summary relies on the entity and proper noun information available in the document, we replace all named entities and proper nouns in the document with `[MASK]` tokens. E.g.:

$$\mathcal{I}_{Y_{\text{ent}}}(x_j) = \begin{cases} \texttt{[MASK]}, & \text{if } x_j \text{ is a proper noun} \\ x_j, & \text{or named entity.} \\ & \text{otherwise} \end{cases}$$

where we adopt the convenient abuse of notation $\mathcal{I}_{Y_{\text{ent}}}(x_j)$ to denote the jth output of $\mathcal{I}_{Y_{\text{ent}}}(x)$.

Shuffle Tokens. Besides entities and proper nouns themselves, faithful summaries require correct resolution of their *relations* and *modifications*. Thus the second set of tokens we consider Y_{rel} is comprised of all of the verbs, adjectives, adverbs, and adpositions in y. To drop the relation and modification information, we randomly shuffle all the tokens in the document, i.e., $\mathcal{I}_{Y_{\text{rel}}}(x) = \sigma(x)$ where σ is a random permutation.

Empty Document. Lastly, for the remaining tokens not covered by the two aforementioned strategies, Y_{other}, we discard all tokens in the document, i.e., $\mathcal{I}_{Y_{\text{other}}}(x) = \emptyset$. Stopwords and punctuation are not included in Y_{other}. With empty documents, the summarizer resembles an unconditional language model. While empty documents have been used in previous work (Xu and Durrett, 2021; Xie et al., 2021), we argue that some spurious correlations might emerge from tokens that imply the document topic (e.g., tokens that usually occur with the topics) and aggressively taking empty documents for measuring the faithfulness level of any token would prevent the exposure of such spurious correlations.

3.3 Using MuFaSSA during Training

We modify loss truncation (Kang and Hashimoto, 2020) to enable MuFaSSA for training summarization models. Loss truncation considers tokens that still yield high training losses after several training epochs as noisy tokens and ignores their training losses.[4] For each sample, the training objective with loss truncation is formally defined as:

[4]The loss truncation method is proposed at the sample level. We follow Goyal et al. (2022) to extend loss truncation to the token level.

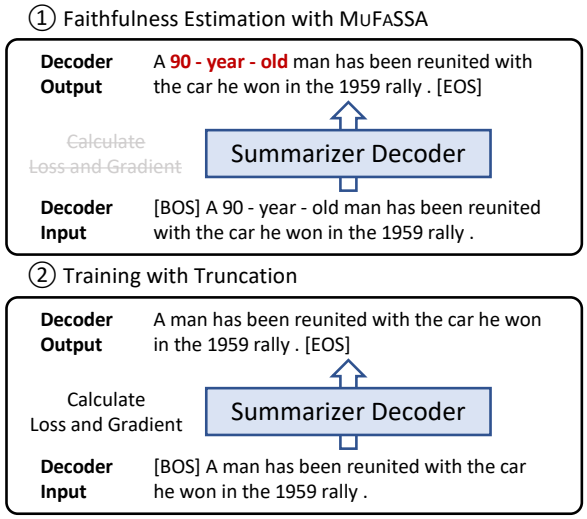

Figure 2: Illustration of our proposed word truncation training objective. In the first pass, model optimization is disabled and MuFaSSA detects less faithful summary tokens. In the second pass, the summarizer is trained on the summary with these tokens discarded.

$$-\frac{1}{L}\sum_{i=1}^{L} \mathbb{1}_{[-\log p_\theta(y_i|y_{<i},x)<Q_c^l]} \cdot \log p_\theta(y_i|y_{<i},x) \tag{2}$$

where Q_c^l is the c percentile of the list Q^l tracking past training losses.

However, high loss might not well estimate the faithfulness level of each token. Thus, we propose to instead use faithfulness scores output by MuFaSSA to identify unfaithful tokens to omit from the loss computation. That is, we replace Q^l with Q^m that records the faithfulness levels of past tokens measured by MuFaSSA. The resulting training objective is:

$$-\frac{1}{L}\sum_{i=1}^{L} \mathbb{1}_{[m(y_i)>Q_c^m]} \cdot \log p_\theta(y_i|y_{<i},x) \tag{3}$$

where the summarizer p_θ that is being optimized is also used for obtaining $m(y_i)$ as in Equation (1). Before switching to our modified training objective, we first optimize the summarization model for several epochs with the traditional cross entropy objective (henceforth, warm-up stage) following Goyal et al. (2022). The number of epochs for the warm-up stage is set to 3 in our experiments. We tune the percentile c on validation sets to achieve a balance of summary faithfulness and coverage.

Word Truncation. Although loss truncation avoids optimizing the likelihood of less faithful

tokens, they are retained in the generation context for the remaining tokens. Thus, the summarization model might insist on generating them in order to generate the remaining content, yielding unfaithful summaries. To this end, we extend loss truncation by additionally removing tokens that are less faithful from the decoder input during training. As illustrated in Figure 2, after feeding the original decoder input to the model, the faithfulness levels of summary tokens are first estimated by MUFASSA in the decoder output. At this step, we do not calculate the loss or perform any gradient back-propagation. With the less faithful tokens detected, we remove them from both the decoder input and output of the training sample. Finally, we train the summarizer with the truncated decoder input and output.

4 Metric Experiments

We first test how well MUFASSA agrees with human judgments on summary faithfulness.

Datasets. We experiment on AGGREFACT (Tang et al., 2022), a benchmark consisting of document-summary pairs and their binary faithfulness labels annotated by most recent work (Kryscinski et al., 2020; Maynez et al., 2020; Huang et al., 2020; Fabbri et al., 2021; Pagnoni et al., 2021; Cao and Wang, 2021; Goyal and Durrett, 2021; Cao et al., 2022). We use the SOTA subset of AGGREFACT where the summaries are produced by state-of-the-art summarizers built from large pre-trained models. The SOTA subset contains 1,335 and 1,018 samples annotated on XSum (Narayan et al., 2018) and CNN/DailyMail (Hermann et al., 2015) respectively.

Comparisons. For comparison, we include results of existing state-of-the-art faithfulness evaluation metrics:

- QUESTEVAL (Scialom et al., 2021) is a QA-based metric that answers questions created from the summary using the document and vice versa. To obtain the evaluation score, the word overlaps between the answers given by the pre-trained QA model and the ground-truth answers used for generating the questions are aggregated over all questions.

- SUMMAC (Laban et al., 2022) is an entailment-based metric that first computes the entailment score between each pair of document and summary sentences. For each summary sentence, its entailment scores with document sentences are then binned into a histogram and transformed into the sentence-level faithfulness score via a 1-D convolutional layer. The mean of the sentence-level scores is then taken as the evaluation score.

- COCO (Xie et al., 2021) is a model causality-based metric. We use its best-performing variant that masks document sentences that contain words in the summary. The difference between summary output probabilities given by a trained summarizer using the original document and the masked document is taken as the evaluation score.

We also compare with two variants of MUFASSA that: (1) directly take the output probability given by the original input as the faithfulness estimation (PROBABILITY), which no longer calculates the difference in Equation 1; or (2) only use the empty document as the ablated input (EMPTY).

Metric	AGGREFACT-XSUM	AGGREFACT-CNN	Average
QUESTEVAL	61.6	**71.5**	66.5
SUMMAC	**66.3**	66.7	66.5
COCO	59.3	68.4	63.8
PROBABILITY	54.7	68.5	61.6
EMPTY	65.1	67.0	66.1
MUFASSA	64.8	69.2	**67.0**

Table 1: The Area Under the ROC Curve (AUC) of different faithfulness metrics on AGGREFACT. The top two results on each split are highlighted with a **boldface** and underline, respectively. MUFASSA achieves better average performance than existing metrics.

Results. The performance by each metric is measured with the Area Under the ROC Curve (AUC). As shown in Table 1, when solely taking the empty document as the ablated input, the resulting metric already matches the performance of the other existing metrics except for QUESTEVAL on CNN/DailyMail, showing the effectiveness of ablation.

Furthermore, boosted by multi-view information ablation that provides model interpretation of finer granularity, MUFASSA *yields the best average performance on* AGGREFACT, even without leveraging models obtained from other datasets.

We also observe that the average performance of COCO is worse than the empty document ablation,

though CoCo employs a more sophisticated masking strategy. As their masking strategy is based on exact word matching, it might struggle to ablate information for abstractive summaries, leading to less accurate faithfulness estimation.

5 Summarization Experiments

To verify the effectiveness of our methods to produce more faithful summaries, we train summarizers on popular summarization datasets with our proposed loss truncation and word truncation methods equipped with MuFaSSA.

5.1 Experimental Setup

Datasets. We conduct experiments on XSum (Narayan et al., 2018) and CNN/DailyMail (Hermann et al., 2015) datasets. Both datasets are built from news articles, with the XSum summaries tending to be more abstractive than its counterpart. We follow the official data splits of XSum and CNN/DailyMail, which respectively contain 204,045/11,332/11,334 and 287,113/13,368/11,490 samples in the train/validation/test sets.

Evaluation Metrics. For faithfulness evaluation, we use SummaC and QuestEval, which respectively obtain the best performance on the XSum and CNN/DailyMail splits of the AggreFact benchmark in §4. In addition, we report ROUGE scores, including variants based on the unigram overlap (ROUGE-1), bigram overlap (ROUGE-2), and longest common subsequence (ROUGE-L) [5].

Comparisons. Besides the models fine-tuned only with the cross entropy objective (BART) and additionally with loss truncation (LossTrunc), we also compare with DAE-based loss truncation (Goyal and Durrett, 2021) and CLIFF (Cao and Wang, 2021). Specifically, DAE assesses the entailment level of each dependency arc in the summary and then locates less faithful tokens by aggregating the entailment levels of their attached dependency arcs, where the training losses are discarded. By contrast, without using truncating losses, CLIFF augments the model training with negative samples (i.e., synthetic incorrect summaries) and adopts contrastive learning (Khosla et al., 2020) to help model distinguish incorrect summaries from correct summaries.

[5] Please refer to Appendix B for ROUGE-1 and ROUGE-2 scores

Model	SummaC	QuestEval	R-L
XSum			
BART	24.36	36.66	**37.19**
CLIFF	24.60*	36.94	36.43
DAE	23.81	36.38	30.32
LossTrunc	24.52	37.12*	34.60
+ MuFaSSA	24.63*	**37.22***	33.77
+ WordTrunc	**24.85***	36.75	34.66
CNN/DailyMail			
BART	80.54	60.17	41.14
CLIFF	78.95	59.03	41.06
LossTrunc	80.50	60.22	**41.36***
+ MuFaSSA	81.84*	60.04	40.69
+ WordTrunc	**83.01***	**60.44***	40.40

Table 2: Evaluation of summary generation on XSum and CNN/DailyMail. R-L: ROUGE-L. MuFaSSA-based loss and word truncation yields summarizers with the best faithfulness scores. *Significantly better than BART with approx. randomization test ($p < 0.005$).

5.2 Results

We report results on XSum and CNN/DailyMail in Table 2. Our modified loss truncation produces summarizers with better performance than all comparisons on faithfulness metrics on both datasets, except for QuestEval on CNN/DailyMail, which suggests the usefulness of MuFaSSA in training summarization models with improved faithfulness. Moreover, additionally truncating the less faithful tokens in the decoding context during training consistently advances the SummaC scores, achieving the best SummaC scores on both datasets.

Though DAE trains a dependency arc entailment scorer with augmented negative samples, the external dependency parser requires processing the summarization dataset into a text format that does not align with the natural text format used by large model pre-training, yielding worse performance.

Additionally, we observe that summaries from our models have less competitive ROUGE scores. This could be due to unfaithful content in the human reference summaries, which has been identified as an issue in previous work (Maynez et al., 2020). In this regard, further human evaluation is conducted in the next section.

Human Evaluation. We hire human annotators on Amazon Mechanical Turk[6] to rate system summaries on three aspects:

[6] https://www.mturk.com/

Model	Faith.	Cover.	Coher.
BART	3.32	3.22	**5.00**
CLIFF	3.45	3.22	4.97
LOSSTRUNC	3.41	3.17	4.97
+ MUFASSA	3.45	3.21	4.97
+ WORDTRUNC	**3.59***	**3.35**	4.92

Table 3: Human evaluation results on XSum. Faith.: faithfulness; Cover.: coverage; Coher.: coherence. Our model using word truncation guided by MUFASSA achieves the best summary faithfulness and coverage. Krippendorff's $\alpha \geq 0.35$ for all aspects.

- **Faithfulness**: How well the factual information in the summary accurately matches the information in the article;
- **Coverage**: How well the summary covers the important information in the article; and
- **Coherence**: How coherent the summary is on its own.

Each aspect is rated on a Likert scale from 1 (worst) to 5 (best).

We randomly select 80 articles from XSum, where models are more prone to errors (Pagnoni et al., 2021), and ask annotators to judge summaries generated by our models as well as comparisons including BART, CLIFF, and the original loss truncation. During annotation, the order of the system summaries is shuffled and each system summary is evaluated by three annotators. Details of the human evaluation such as payment, annotator qualification, and interface screenshots are included in Appendix C.

According to human judges (Table 3), without word truncation, MUFASSA improves the identification of less faithful tokens, outperforming the original loss truncation and matching CLIFF on summary faithfulness and coverage. Adding word truncation further encourages the summarizer to generate summaries with promoted faithfulness and content coverage, leading to the best scores on both aspects. We also find that removing less faithful summary tokens from the training samples only has minor effects on the summary coherence.

Case Study. Figure 3 displays an example article from XSum and its corresponding summaries generated by summarizers trained with different methods. The model trained with the original loss truncation does not attempt to modify the unfaithful entity "the Six Nations", as training losses do not accurately reflect faithfulness levels. While the unfaithful entity is removed from the output when the original loss truncation is augmented with MUFASSA, the summarizer produces another piece of unfaithful information. After applying word truncation, the model learns to stop generation, producing the faithful summary.

Article: Amos dislocated a shoulder in the 32-8 defeat by Australia and will have an operation in the next week. The 22-year-old Dragons wing tweeted: "Operation set for Monday, aiming to be back in February". "It's unlucky for Hallam but a great opportunity for Keelan," said Wales assistant coach Neil Jenkins ... "We're going to miss him, but back-three is a position where we have strength in depth." Giles has been in outstanding form for Ospreys, scoring eight tries in five appearances for the region this season ...

BART: Ospreys wing Keelan Giles could make his Wales debut after Hallam Amos was ruled out of the Six Nations with a shoulder injury.

LOSSTRUNC: Ospreys wing Keelan Giles has been named in Wales' back-three after Hallam Amos was ruled out of the Six Nations.

LOSSTRUNC + MUFASSA: Ospreys wing Keelan Giles could make his Wales debut after Hallam Amos was ruled out for the rest of the season.

LOSSTRUNC + MUFASSA + WORDTRUNC: Ospreys wing Keelan Giles is in line to replace injured Hallam Amos in Wales' back-three.

Figure 3: Example generated summaries. Unfaithful information is shaded with red. Our model trained with word truncation signaled by MUFASSA generates a faithful summary.

6 Additional Experiments

In this section, we inspect the effects of important design choices in MUFASSA (§6.1). Furthermore, to show the possibility of applying MUFASSA to other tasks, we train data-to-text generation models with our proposed methods (§6.2).

6.1 Ablation Study

We examine the effects on faithfulness estimation by the source ablations with masked entities and proper nouns, and shuffled tokens. For the faithfulness levels of summary tokens induced by each ablated input, when the ablated input is not used,

Metric	AGGREFACT-XSUM	AGGREFACT-CNN	Average
MUFASSA	64.8	**71.2**	**68.0**
Not Using All Source Ablations			
w/o Mask	64.5	69.8	67.1
w/o Shuffle	**65.3**	69.9	67.6
w/o Mask & Shuffle	65.1	67.0	66.1
Not Assigning Ablations to Different Summary Tokens			
Average	62.4	67.6	65.0
Minimum	53.5	51.3	52.4
Maximum	65.2	68.2	66.7

Table 4: The Area Under the ROC Curve (AUC) by variants of MUFASSA on AGGREFACT. The best results on each split is highlighted with **boldface**. Removing any component of MUFASSA reduces its robustness, leading to lower average AUC.

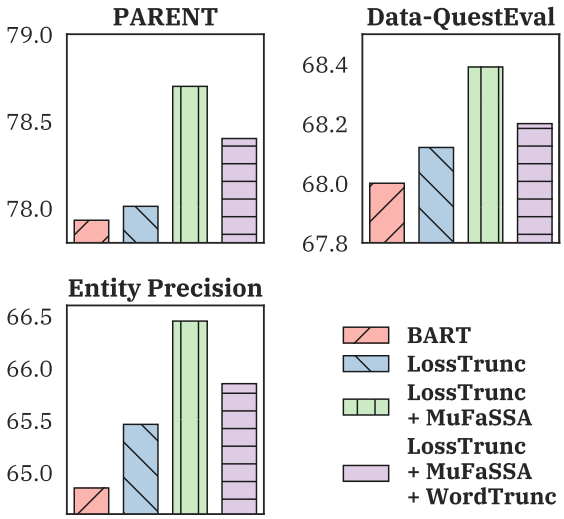

Figure 4: Automatic evaluation results on WikiPerson. Our models achieve better performance than comparisons with the cross entropy objective and original loss truncation objective, implying the effectiveness of MUFASSA on other generation tasks.

we instead obtain their faithfulness levels with the empty document input. Moreover, we investigate the benefits of assigning each ablated input to different summary tokens. Concretely, we consider three variants, where the faithfulness level of each token is calculated by either taking the average, minimum, or maximum value of the faithfulness levels measured with the three source ablations.

Including multiple source ablations enhances the robustness of MUFASSA, as indicated by its best average performance on AGGREFACT in Table 4. Compared to the source ablation with shuffled tokens, the source ablation with masked entities and proper nouns contributes more to the accurate faithfulness estimation by MUFASSA, dropping which leads to a larger performance degradation.

Simple aggregations (i.e., average, minimum, and maximum) of the faithfulness levels measured by the three source ablations produce lower AUC scores, justifying MUFASSA's design of leveraging different token-specific source ablations.

6.2 Extension to Data-to-Text Generation

While this work focuses on summarization, we also explore extending our methods to other tasks. Specifically, we conduct experiments on a data-to-text dataset, WikiPerson (Wang et al., 2018) which requires the generation model to produce a natural language description for a person's career, given the infobox in the corresponding Wikipedia biography article. Details of the dataset and experiment setup are included in Appendix A.2.

We evaluate outputs with faithfulness-aware data-to-text metrics, including: PARENT (Dhingra et al., 2019) that additionally aligns n-grams from the reference and the system generation to the source table; and Data-QuestEval (Rebuffel et al., 2021) which replaces the text-based question generation and answering models in the original QUESTEVAL with table-based models to adapt to data-to-text tasks. Moreover, we compute the precision of named entities in the generated text, suggested by recent work on text generation (Logan IV et al., 2022).

Our models outperform comparisons on all metrics, as shown in Figure 4, indicating the potential adaptations of MUFASSA on conditional generation tasks other than text summarization to improve output faithfulness. Word truncation does not further improve the performance on WikiPerson. We suspect that data-to-text tasks might require more samples to learn coherent generation due to the modality difference between the input and output, while word truncation reduces the number of tokens that the model can learn from.

7 Conclusion

We studied improving faithful summary evaluation and generation. Our proposed method, MUFASSA, estimates the faithfulness level of a summary token by the decrease in its generation probability after ablating crucial information from the source document. Multiple ablation strategies are used by MUFASSA for different summary tokens to achieve accurate faithfulness estimation. We also

designed word truncation for improved integration of MUFASSA into model training. Experiments on AGGREFACT show that MUFASSA better aligns with human faithfulness labels than existing metrics. When used for highlighting less faithful tokens during summarizer training, MUFASSA leads to summaries with enhanced faithfulness, which is further boosted by word truncation, achieving better faithfulness than competitive comparisons, as measured by both automatic metrics and human annotators.

Limitations

While MUFASSA does not rely on textual entailment or question answering models, the construction of ablated inputs in MUFASSA still requires some existing NLP tools such as named entity recognizers and POS taggers. Therefore, the accuracy of the faithfulness estimation would be limited by the performance of these tools. Also, construction strategies other than the ones presented in this paper might rely on more advanced NLP tools, further amplifying the limitation. This could be a significant issue for low-resource languages where basic NLP tools have not been established.

In addition, our word truncation training objective incurs some computational overhead. First, it takes two forward passes, though gradient backpropagation is not performed in the first pass. Second, similar to the original loss truncation, word truncation maintains a list for storing the faithfulness levels of past tokens and needs to calculate the threshold of faithfulness levels for truncating less faithful tokens.

Ethical Consideration

Previous studies have shown that large pre-trained models embed biases and might create harm to certain populations. While MUFASSA is built with large pre-trained models, we do not study if the faithfulness estimation by MUFASSA is biased towards any population in this work (e.g., produce higher scores for texts including a population than text including another population). As recent work finds that BERTScore which is also based on large pre-trained models has biases (Sun et al., 2022), we suggest users carefully investigate the potential biases in the model before applying it in real-world situations.

References

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

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *thirty-second AAAI conference on artificial intelligence*.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. 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.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. 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.

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.

Tanya Goyal, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. 2022. Training dynamics for text summarization models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2061–2073, Dublin, Ireland. 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.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

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.

Daniel Kang and Tatsunori B. 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. *Advances in Neural Information Processing Systems*, 33:18661–18673.

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. 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.

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.

Wei Liu, Huanqin Wu, Wenjing Mu, Zhen Li, Tao Chen, and Dan Nie. 2021. Co2sum: Contrastive learning for factual-consistent abstractive summarization. *arXiv preprint arXiv:2112.01147*.

Robert L. Logan IV, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2022. FRUIT: Faithfully reflecting updated information in text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3670–3686, Seattle, United States. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. 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.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. Bertscore is unfair: On social bias in language model-based metrics for text generation.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, 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.

Qingyun Wang, Xiaoman Pan, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji, and Kevin Knight. 2018. Describing a knowledge base. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 10–21, Tilburg University, The Netherlands. Association for Computational Linguistics.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiacheng Xu and Greg Durrett. 2021. Dissecting generation modes for abstractive summarization models via ablation and attribution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6925–6940, Online. Association for Computational Linguistics.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

Original Paper	AGGREFACT-XSum	AGGREFACT-CNN
Polytope (Huang et al., 2020)	-	68
SummEval (Fabbri et al., 2021)	-	400
FRANK (Pagnoni et al., 2021)	-	250
Wang'20 (Wang et al., 2020)	239	-
CLIFF (Cao and Wang, 2021)	300	300
Goyal'21 (Goyal and Durrett, 2021)	100	-
Cao'22 (Cao et al., 2022)	696	-

Table 5: Numbers of samples collected from previous work in the SOTA subset of AGGREFACT.

Model	R-1	R-2	Density	Coverage
XSum				
BART	**45.41**	**22.29**	1.65	75.70
CLIFF	44.52	21.40	1.69	76.71
DAE	38.94	15.00	1.50	74.80
LOSSTRUNC	42.98	19.13	1.74	78.78
+ MUFASSA	41.93	18.09	1.85	77.95
+ WORDTRUNC	42.36	19.13	1.75	76.82
CNN/DailyMail				
BART	44.32	21.32	20.81	99.00
CLIFF	44.18	21.14	18.89	98.91
LOSSTRUNC	**44.50**	**21.48**	20.16	99.02
+ MUFASSA	43.88	20.96	21.52	99.15
+ WORDTRUNC	43.63	20.77	24.57	99.32

Table 6: ROUGE-1 and ROUGE-2 on XSum and CNN/DailyMail. The best result of each metric on each dataset is **bolded**.

A Details of Datasets

We include additional details for datasets we use in our paper.

A.1 AGGREFACT

We show the numbers of samples included in the SOTA subset of AGGREFACT (Tang et al., 2022) from different studies in Table 5.

A.2 WikiPerson

WikiPerson (Wang et al., 2018) extract Wikipedia articles and the corresponding infoboxes about person entities. For each article, they remove sentences that do not contain any value in the corresponding infobox or only contain entities not in the infobox. The remaining sentences of the article are then taken as the generation target for the infobox.

Statistics. We use the official data split provided by the original paper, which contains 250,186/30,487/29,982 samples in the train/validation/test sets. On average, each infobox contains 7.3 attribute-value pairs and each target output contains 86.3 words.

Experiment Details. On WikiPerson, MUFASSA masks the values in the infoboxes for estimating the faithfulness levels of entities and proper nouns in the outputs. For the remaining tokens in the outputs, we use empty infoboxes as the ablated inputs. We do not consider shuffling tokens of values in infoboxes, as they are mainly entities and proper nouns.

Given an infobox, to create the textual input to the model, we first concatenate attributes and their corresponding values with "=". Then we concatenate all attribute-value pairs together with "|" inserted at the beginning of each attribute-value pair. An example of the converted textual input: "| *Name_ID = Thorsten Barg | date of birth = 25 August 1986 | country of citizenship = Germany*".

B Additional Results

We report ROUGE-1 and ROUGE-2 scores on XSum and CNN/DailyMail, which are omitted in §5.2. Both scores follow the trend of ROUGE-L in Table 2. We also examine the abstractiveness of the summaries generated by each system by calculating the density and coverage (Grusky et al., 2018), where we find our system tends to be more extractive on CNN/DailyMail compared to other systems.

C Details of Human Evaluation

Our human evaluation is conducted on Amazon Mechanical Turk (AMT). In the annotation interface (Figure 5 to 8), we provide a detailed instruction of the annotation task, including rubrics, examples, and explanations.

Before launching all annotation samples on AMT, we run two batches for qualification. Each qualification batch contains one article and its corresponding system summaries, and is annotated by 20 workers. We manually inspect the annotation results and filter out workers that return abnormal annotations (e.g., giving high faithfulness scores to

for summaries containing unfaithful content or giving very different scores to identical summaries). We also require the annotators to be located in the US or the UK, with 100 tasks previously completed, and have an approval rate of 100%. A pool of 8 workers for our human evaluation is obtained after the qualification.

For compensation, we pay each annotator $2.5 for each task (i.e., evaluating system summaries generated for an article) of our human evaluation to approximate an hourly payment of $15.

D Details of Implementation

We use Fairseq (Ott et al., 2019)[7] for setting up the training and decoding pipelines. The BART model (Lewis et al., 2020) in our paper is initialized from the `bart.large`[8] checkpoint provided by Fairseq. We conduct training and decoding on 4 NVIDIA V100 GPUs with 16GB memory.

Training. We use the training hyperparameters in the training script provided by the BART paper[9]. The percentile for obtaining the threshold of faithfulness levels is tuned on the validation set of each dataset. For XSum, we search for the best threshold percentile within $[30, 40, 50]$. The model with the best SUMMAC score while having a ROUGE-1 score of at least 42 is selected. 40, 50, and 30 are chosen as the percentiles for the models respectively trained with the original loss truncation objective, our loss truncation guided with MUFASSA, and our word truncation objective. For CNN/DailyMail, we search for the best threshold percentile within $[5, 10, 20]$. The model with the best SUMMAC score while having a ROUGE-1 score of at least 44 is selected. 10, 10, and 5 are chosen as the percentiles for the models respectively trained with the original loss truncation objective, our loss truncation guided with MUFASSA, and our word truncation objective. To avoid incoherent summaries, we only apply word truncation to proper nouns. Due to the computational cost, we train all models for one run.

Decoding. We follow the original BART paper and decode using beam search with beam sizes of 4 and 6 on CNN/DailyMail and XSum. During decoding, the maximum decoding lengths are 140 and 60 for CNN/DailyMail and XSum.

Running Time. We report the running time on XSum. Training our models with loss truncation or word truncation on XSum takes 10 hours and the decoding takes half an hour.

Model Parameters. Our methods do not increase the number of model parameters. Therefore, our models have 400M parameters, which is the same as the original BART.

E Output Examples

We include more examples of system outputs in Figure 9 and 10.

[7] https://github.com/pytorch/fairseq

[8] https://github.com/pytorch/fairseq/tree/main/examples/bart

[9] https://github.com/pytorch/fairseq/blob/main/examples/bart/README.summarization.md

Task Instructions

There will be many similar HITs for you to perform if you do well at this task. Please follow the instructions carefully for each HIT; we will be reviewing your HITs periodically and if we note any unusual responses, you might not see any additional tasks from us.

During this task, you will read a news article and six different summaries for the article. You will rate the quality of each of the six summaries by four axes: *coherence*, *accuracy*, *coverage*, and *overall quality*.

The rubrics below give specific guidance on how each axis should be rated. Please read the rubrics carefully before continuing to the task.

Jump to coherence rating

Coherence

For each summary, answer the question *"how coherent is the summary on its own?"* (on a scale from 1 to 5). A summary is *coherent* if, when read by itself, it's easy to understand and free of English errors. A summary is not coherent if it's difficult to understand what the summary is trying to say. Generally, it's more important that the summary is understandable than it being free of grammar errors. Please **do not penalize incomplete punctuation** (e.g., when there exists only one quote mark in the sentence).

Rubric:

- Score of 1: The summary is impossible to understand.
- Score of 2: The summary has many mistakes or confusing phrasing.
- Score of 3: The summary has some mistakes or confusing phrasing that make it hard to understand.
- Score of 4: The summary has only one or two mistakes or confusing phrasing.
- Score of 5: The summary is perfectly clear.

Jump to accuracy rating

Accuracy

For each summary, answer the question "how well does the factual information in the summary accurately match the information in the article?" (on a scale of 1 to 5) A summary is accurate if it doesn't say things that aren't in the article, it doesn't contradict information in the article, and generally is not misleading.
Even if a piece of information is true according to your knowledge, if it is not mentioned in the article it should not be included in the summary.

Rubric:

- Score of 1: The summary is completely wrong, made up, or exactly contradicts what is written in the article.
- Score of 2: The summary says many things not mentioned in or contradicting the article.
- Score of 3: The summary says at least one substantial thing that is not mentioned in the article, or that contradicts something in the article.
- Score of 4: The summary says anything at all that is not mentioned in the article or contradicts something in the article.
- Score of 5: The summary has no incorrect statements or misleading implications.

Jump to coverage rating

Coverage

For each summary, answer the question *"how well does the summary cover the important information in the article?"* (on a scale of 1 to 5). A summary has good *coverage* if it mentions the main information from the article that's important to understand the event described in the article. A summary has poor coverage if someone reading only the summary would be missing several important pieces of information about the event in the article.

Rubric:

- Score of 1: The summary contains no information relevant to the article.
- Score of 2: The summary is missing many important pieces of information required to understand the event.
- Score of 3: The summary is missing at least one crucial piece of information required to understand the event.
- Score of 4: The summary is missing any information (no matter how small) required to understand the event.
- Score of 5: The summary covers all of the important information required to understand the event.

Jump to overall quality rating

Overall quality

For each summary, answer the question *"how good is the summary overall at representing the article?"* (on a scale of 1 to 5). This encompasses all of the above axes, as well as the information included in the summary and if it has helped you understand the event. If it's hard to find ways to make the summary better, give the summary a high score. If there are lots of different ways the summary can be made better, give the summary a low score.

Figure 5: Screenshot of our annotation interface (1/4).

Rubric:

- Score of 1: The summary is terrible.
- Score of 2: The summary is a pretty bad representation of the article and needs significant improvement.
- Score of 3: The summary is an okay representation of the article, but could be significantly improved.
- Score of 4: The summary is a pretty good representation of the article, but it's not perfect.
- Score of 5: The summary is an excellent representation of the article.

Example

Now you will review an example article and three associated summaries.

- For each of the summaries, we have provided ratings on the four axes: coherence, accuracy, coverage, and overall quality with explanations for why those ratings were chosen.
- Please review the summaries and their ratings carefully, so you understand how to rate the summaries during the task.
- If you have any questions about how to rate the summaries, please consult the rubric (above).

Example Article

Welsh and UK ministers have been rowing since March over how to finance the commuter lines in and out of Cardiff. Mr Crabb said the scheme - estimated at £309m to £463m - was "probably the most knotty" problem between the two governments but was soluble. The valleys rail electrification is due to be completed between 2019 and 2024. Planned rail improvements will see the upgrade of the main line from London Paddington to Cardiff, which is due to be completed by 2017, and extended to Swansea by 2018 at a cost of £850m. The electrification of the Valleys lines was due to follow, but the plan was thrown into doubt in March by a row over the financing of the project. Speaking on Radio Wales' Sunday Supplement programme, Mr Crabb said rail electrification was the "number one issue" for him. He said: "It's something that I've been spending quite a bit of my summer working on. "There's a bit more work to be done between the two governments on where we think the solution lies, but I think when I go around talking to businesses in south Wales they are desperate to see this problem answered, they want the two governments to be working effectively together." Describing the issue as "a bit of a litmus test" for joint working between Wales and Westminster, he warned the issue "can't drag on indefinitely". "There are engineering teams involved in Network Rail who need to get tasks assigned to them if this huge, enormous, financially-challenging project is to go ahead," he said. "There are some quite hard deadlines in that. But we are talking a short number of months hopefully."

Example summary #1

The electrification of the Valleys rail lines is the "number one issue" for Welsh Secretary Stephen Crabb.

Ratings for Example Summary #1

The summary should be rated as follows:

How coherent is the summary on its own?

It is impossible to understand. 1 2 3 4 5 The summary is perfectly clear.

Explanation for rating: This summary is easy to understand and read with clear language and no grammatical errors and therefore coherent, so we rate it a 5 (of 5).

How well does the factual information in the summary accurately match the article?

The summary is completely wrong, made up, or exactly contradicts what is written in the article. 1 2 3 4 5 The summary has no incorrect statements or misleading implications.

Explanation for rating: The position and first name of Mr. Crabb is unknown from the article. So we rate this summary as 3 (of 5).

How well does the summary cover the important information in the article?

The summary contains no information relevant to the article. 1 2 3 4 5 The summary covers all of the important information required to understand the event in the article.

Explanation for rating: This summary has a fair coverage of the article, but it misses the mention of the underlying reason for the rail line electrification issue. So we rate this summary as 4 (of 5).

How good is the summary overall at representing the article?

It is terrible. 1 2 3 4 5 It is an excellent representation of the article.

Explanation for rating: This summary is okay but it could be **significantly improved** by mentioning the underlying reason for the rail line electrification issue and not including extraneous information about Mr. Crabb. So, we rate this summary as 3 (of 5).

Example Summary #2

Mr. Crabb has said he is "desperate" to see the electrification of the Valleys rail line.

Ratings for Example Summary #2

The summary should be rated as follows:

How coherent is the summary on its own?

It is impossible to understand. 1 2 3 4 5 The summary is perfectly clear.

Explanation for rating: This summary is easy to understand and read with clear language and no grammatical errors and therefore

Figure 6: Screenshot of our annotation interface (2/4).

coherent, so we rate it a 5 (of 5).

How well does the factual information in the summary accurately match the article?

The summary is completely wrong, made up, or exactly contradicts what is written in the article. 1 2 3 4 5 The summary has no incorrect statements or misleading implications.

Explanation for rating: It could be inferred that Mr. Crabb is "desperate", but it is not explicitly stated in the article. So we rate this summary as 4 (of 5).

How well does the summary cover the important information in the article?

The summary contains no information relevant to the article. 1 2 3 4 5 The summary covers all of the important information required to understand the event in the article.

Explanation for rating: This summary has a fair coverage of the article, but it misses the mention of the underlying reason for the rail line electrification issue. So we rate this summary as 4 (of 5).

How good is the summary overall at representing the article?

It is terrible. 1 2 3 4 5 It is an excellent representation of the article.

Explanation for rating: This summary is pretty good, but it can be **somewhat improved** by providing the underlying reason for the rail line electrification issue. So we rate this summary as 4 (of 5).

Example Summary #3

The electrification of the Valleys rail lines interrupted by the finance plan is the "number one issue" for Crabb.

Ratings for Example Summary #2

The summary should be rated as follows:

How coherent is the summary on its own?

It is impossible to understand. 1 2 3 4 5 The summary is perfectly clear.

Explanation for rating: This summary is easy to understand and read with clear language and no grammatical errors and therefore coherent, so we rate it a 5 (of 5).

How well does the factual information in the summary accurately match the article?

The summary is completely wrong, made up, or exactly contradicts what is written in the article. 1 2 3 4 5 The summary has no incorrect statements or misleading implications.

Explanation for rating: Information in this summary is accurately grounded in the article. So we rate this summary as 5 (of 5).

How well does the summary cover the important information in the article?

The summary contains no information relevant to the article. 1 2 3 4 5 The summary covers all of the important information required to understand the event in the article.

Explanation for rating: This summary contains all important information in the article. So we rate this summary as 5 (of 5).

How good is the summary overall at representing the article?

It is terrible. 1 2 3 4 5 It is an excellent representation of the article.

Explanation for rating: This summary is an excellent representation of the article. So we rate this summary as 5 (of 5).

Test

Answer the following question to start the task. If you are unsure of the answer, review the rubrics above. The task section will appear when you've completed the test.

Which axis measures whether the summary information is grounded in the article information? Enter your answer and click "Start task".

accuracy Correct!

Task

Instructions:

1. Read the article and when you are finished reading, click "Yes".
2. Write a short title for the article then click "Submit title".

Figure 7: Screenshot of our annotation interface (3/4).

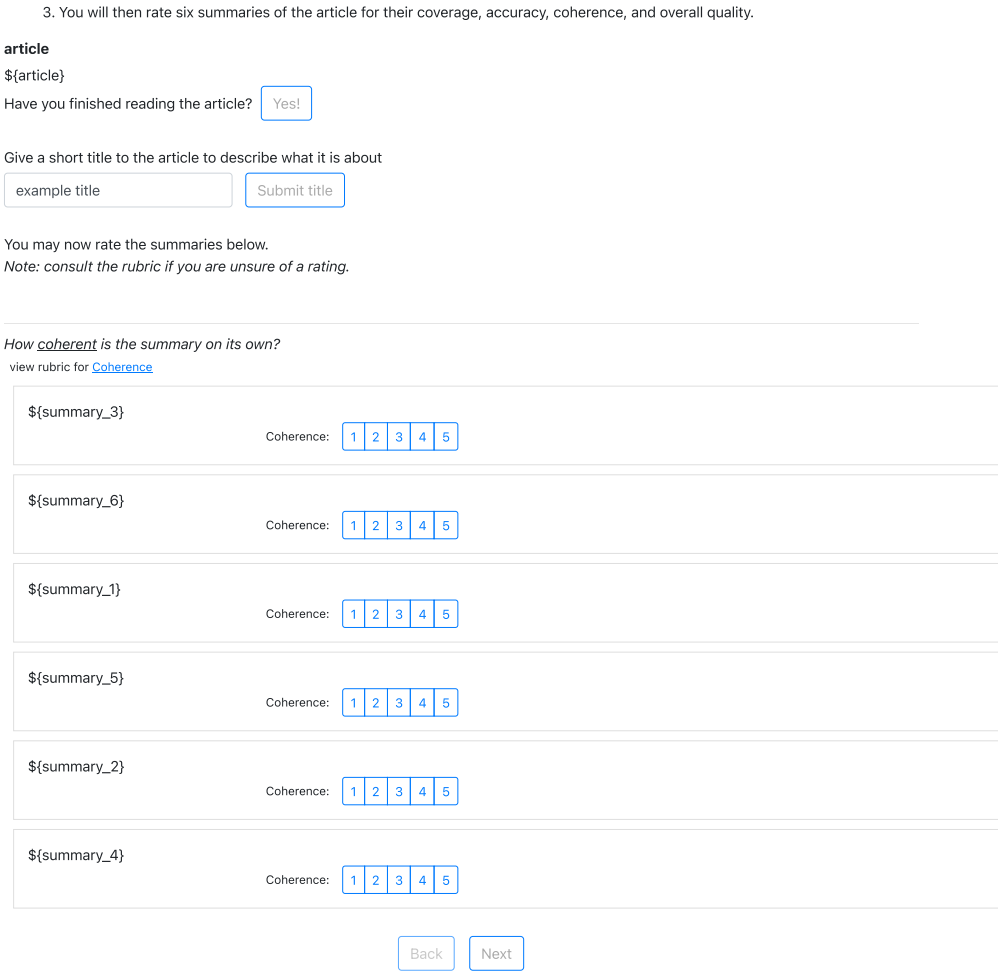

Figure 8: Screenshot of our annotation interface (4/4).

Article: Downing Street backed a report by think tank Policy Exchange which said selling high value homes when they become vacant would raise £4.5bn a year. That would be enough to build 80,000 to 170,000 social homes, the report said. Labour said new homes were urgently needed but "driving out hard-working families on low wages from whole neighbourhoods" was not the answer. In its Ending Expensive Social Tenancies report, Policy Exchange argues the move could create the largest social house building programme since the 1970s - giving the economy a kickstart. Neil O'Brien, the think tank's director, told the BBC that social housing would still exist in very expensive areas under its proposal, but there would just be "less of it". "The truth is I don't believe anybody has the right to live in the most expensive parts of town. "People do have a right to get housed, just not in the very most expensive areas," he said. He also suggested that the overall number of people waiting for social housing, currently around 1.8 million, could be reduced by about 500,000 if the scheme was implemented. The prime minister's official spokesman said: "This is something that councils can choose to do already. "Councils should be looking for ways to use their social housing stock as efficiently as they can. The waiting list for social housing has increased a lot over passing years. "They need to think about how they can use that social housing stock efficiently. "If they can sell high-value housing to invest in more social housing and find more homes for more people, then that is certainly something they should look at." But Labour said the coalition's "failed" polices were "making the housing crisis worse not better". Shadow housing minister Jack Dromey said: "Councils and housing associations should make effective use of their housing stock but the government should not force them to arbitrarily sell off social homes, breaking up mixed communities and driving out hard-working families on low wages from whole neighbourhoods." He said the government should use a bank bonus tax to fund 250,000 affordable homes and "put unemployed builders back to work" and boost the construction industry. 'Lucky family' Expensive social housing - which Policy Exchange defines as housing worth more than the average property in each region - accounts for 21.8% of the total social housing stock in the UK, it says. This equates to 816,000 properties - out of a total of 3.78 million - which the think tank says could raise up to £159bn if sold. It says London alone has more than £70bn of expensive social housing. About 3.5% of the total stock becomes vacant every year owing to people moving out or dying, the think tank said. This meant the government could sell a total of 28,500 properties each year, raising £5.5bn a year. The figure would stand at £4.5bn after paying off the debt held against the stock, the report said. Mr O'Brien argued that many hard-working people might want to live in a nicer area or in a bigger house but could not afford to. "Rather than having one lucky family with a very expensive house, you would have two families perhaps desperately waiting for social housing, now having a roof over their heads. "That seems fairer to me," he added. The think tank also said the move would be "extremely popular" with all sections of society, claiming that 73% of people, including social tenants, think people should not be given council houses worth more than the average property in a local authority. 'Dramatic erosion' Critics say such a move would push the least well-off out of expensive streets, and into new ghettos. The National Housing Federation, which represents housing associations, says many towns would be "cleansed" of "hardworking people who can't afford to pay high prices". Labour MP Karen Buck, who represents Westminster North, is concerned that lower income families, particularly in London, will be forced out of more affluent areas creating segregated communities of rich and poor. Ms Buck also argued that the Labour government's £8bn social and affordable housing building programme was cut by 60% when the coalition came to power. Housing Minister Grant Shapps - who is in favour of a sell-off - said the government had introduced "radical reforms" to "get Britain building" and to reduce social housing waiting lists. They included investing £19.5bn public and private funding into an affordable housing programme "set to exceed expectations and deliver up to 170,000 homes". Councils could now offer fixed-term tenancies to new tenants to make sure "social housing goes to those in greatest need", he added.

BART: Prime Minister David Cameron has said councils should be allowed to sell off expensive social housing to fund more affordable homes.

LOSSTRUNC: The government has said it would be "appropriate" for councils to sell off social housing in very expensive areas.

LOSSTRUNC + MUFASSA: Councils should sell high-value social housing to help build more homes, the prime minister's office says.

LOSSTRUNC + MUFASSA + WORDTRUNC: Councils should be allowed to sell council houses worth more than the average property to fund new homes, the government says.

Figure 9: Output examples on XSum.

Article: Following Raheem Sterling's interview on Wednesday, in which he said he was not ready to sign a new contract at Liverpool, blogger David Tyrer of Live4Liverpool gives the view from Merseyside. While I hate to use social media as a gauge of opinions, Raheem Sterling's interview didn't go down well at all. It was ill-timed and, regardless of what Sterling and his agent hoped, he didn't come across very well. Some of his answers only fuelled the fire really. I'm hoping that that wasn't the whole point, as we've seen these sorts of situations engineered by agents before. The interview has almost certainly changed the way the fans feel about him. There will be a lot of fans that are of the opinion: 'let him go'. Obviously, with the caveat that we get our money's worth! Raheem Sterling returns to Liverpool training after the international break and shakes hands with manager Brendan Rodgers . Sterling risked angering Liverpool fans after he said in an interview he was not ready to sign a contract . Sterling trains ahead of the weekend's game with Arsenal . It's always disappointing when a young player gets his head turned, but there's a sense of ungratefulness about the whole situation, considering how the club has nurtured him and paid him well throughout. Personally, I think he has the potential to be worth so much more than the £100,000-a-week contract he's turned down. But it's only that: potential. At present, he's arguably in the top five best young players in the world but, obviously at his age, he's also prone to bouts of inconsistency and prolonged poor form. He hasn't been great recently and was awful against Man Utd. Sterling has been linked with a move to Arsenal - the team he is preparing to face at the weekend . Raheem Sterling played for England in the 4-0 Euro 2016 qualifier against Lithuania . And while I'd be willing to see the club give him £100k a week – possibly £120k a week - the club shouldn't do everything it can to keep him. Definitely not. Liverpool fans have a popular mantra: no player is bigger than the club. Admittedly, we stretch the rules for truly great players (Suarez, a recent example) but Sterling is nowhere near. If he wants out, I'm sure the club will handle it the way they have before. Frustration over Sterling's situation has been building for a while, and many fans are now of the belief that if he wants to go he's welcome to. Personally, I don't like players holding the club to ransom. He has as much chance of winning trophies here as he does anywhere (other than money-rich clubs such as Chelsea or Man City). Sterling did not have his best game in a Liverpool shirt during the defeat at Manchester United . Liverpool fans have a mantra that no player is bigger than the club . If Sterling was to move to Arsenal then it would not go down well with Liverpool fans . But if he does end up going, he's worth a lot more than many established players. His ability is so raw but he has almost limitless potential. At the moment, we've seen it in fits and starts but even so, he's easily worth £25-30 million. If I was FSG, I'd hold any interested club to ransom, though – double it. But a move to Arsenal wouldn't go down well. Perhaps better than if it was Chelsea or Manchester United, but honestly there's little Arsenal can offer that we can't. Sure, they're more financially stable but what – in terms of trophies – do Arsenal have to show for the last 12 years? Sterling speaks to the BBC and reveals he is not ready to sign a new contract . I think we're somewhat ahead on that front. Much like Alexis Sanchez, any move to Arsenal will be more about location than anything else. People have drawn comparisons with when Suarez wanted to leave for Arsenal, but the situation is different. Suarez says he wanted Champions League football and we weren't offering that at that time. Not only that, but Suarez was an established top-class player at the time, one of the best in the world. Sterling is purely potential and, given all the club have done for him (cliched, as that sounds), it makes little sense. If he's going to be nurtured into the world-class talent he can become, it won't happen at Arsenal. It may not necessarily happen at Anfield either, but a manager like Brendan Rodgers will give him the best shot of improving to that level over the next two to three years. You can read more from Live4Liverpool HERE and follow the Twitter account HERE .

BART: Raheem Sterling said he was not ready to sign a new contract at Liverpool . The England winger has been linked with a move to Arsenal . Liverpool fans have a mantra that no player is bigger than the club . But a move to Arsenal would not go down well with the Anfield faithful .

LossTrunc: Raheem Sterling said he was not ready to sign a new contract at Liverpool . The England winger has been linked with a move to Arsenal . Liverpool fans have a mantra that no player is bigger than the club . Sterling has the potential to be worth more than the £100,000-a-week contract .

LossTrunc + MuFaSSA: Raheem Sterling has said he is not ready to sign a new contract at Liverpool . The England winger has been linked with a move to Arsenal . Liverpool fans have a mantra that no player is bigger than the club . But a move to Arsenal would not go down well with the fans .

LossTrunc + MuFaSSA + WordTrunc: Raheem Sterling said he was not ready to sign a new contract at Liverpool . Sterling has been linked with a move to Arsenal - the team he is preparing to face at the weekend . Liverpool fans have a mantra that no player is bigger than the club . If Sterling was to move to Arsenal then it would not go down well with Liverpool fans .

Figure 10: Output examples on CNN/DailyMail.