Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors

Anonymous ACL submission

Abstract

The propensity of abstractive summarization models to make factual errors has been the subject of significant study, including work on metrics to detect factual errors and annotation of errors in current systems' outputs. However, the ever-evolving nature of summarization systems, metrics, and annotated benchmarks makes factuality evaluation a moving target, and drawing clear comparisons among metrics has become increasingly difficult. In this work, we aggregate summary factuality error annotations from across nine existing datasets and stratify them according to the underlying summarization model annotated to understand metric performance in scoring state-of-the-art and prior models. To support finer-grained analysis, we unify error types into a single taxonomy based on the function of error word(s) and automatically project each of the datasets' errors into this shared labeled space. We then contrast five state-of-the-art factuality metrics on this benchmark. Our findings show that metric results on datasets built on pretrained model outputs show significantly different results than on datasets with pre-Transformer models. Furthermore, no one metric is superior in all settings or for all error types, and we provide recommendations for best practices given these insights.[1]

1 Introduction

Although abstractive summarization systems (Liu and Lapata, 2019; Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020) have improved dramatically in recent years, these models still often include factual errors in generated summaries (Kryscinski et al., 2020; Maynez et al., 2020). A number of metrics have emerged to detect factuality errors, including methods based on sentence entailment (Kryscinski et al., 2020), finer-grained entailment (Goyal and Durrett, 2020; Zhao et al., 2020), question generation and answering (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021), and discrimination of synthetically-constructed error instances (Cao and Wang, 2021). Despite recent analyses (Pagnoni et al., 2021; Laban et al., 2022), reliably comparing these metrics remains difficult.

To facilitate a careful comparison of factuality metrics, we mainly answer two questions in this paper. First, while current state-of-the-art (SOTA) factuality metrics have made progress in detecting factual inconsistency from summaries, **can these metrics perform well in identifying errors from *state-of-the-art* summarization models (Section 3)?** To answer this question, we create a new benchmark AGGREFACT that consists of nine existing annotated summarization datasets with output from diverse base summarization models ranging from less recent to SOTA ones. We divide our benchmark into three categories SOTA, XFORMER, and OLD based on when the summarization models were developed (Section 2) and compare the performance of factuality metrics across these three categories. We show that current factuality metrics achieve better performance at identifying errors generated by older summarization models. On summaries generated by SOTA models, there is no single metric that is superior in evaluating summaries from both the CNN/DM (Hermann et al., 2015) and XSum (Narayan et al., 2018) datasets.

Second, **what error types are factuality metrics capable of identifying (Section 4)?** We answer this question by leveraging several datasets from our benchmark that have fine-grained annotations. Specifically, we unify error types of these datasets into a single taxonomy for a cross-dataset analysis. We find that the error type distribution changes over time and even differs between annotations of the same summarization models across factuality datasets. Analysis of the factuality metrics shows that metrics claiming SOTA performance can identify each error type better in general, but all metrics differ significantly in how they perform on

[1] Data and code is attached to the submission.

the same error types across CNN/DM and XSum.

We conclude with recommendations for best practices in this area:

1. Prefer evaluating factuality metrics on summaries generated by the state-of-the-art summarization models.

2. Choose an appropriate factuality metric for evaluation at any downstream task at hand. No one metric is superior across all settings.

3. Annotate error types consistently with prior work for better comparability. We found that error type boundaries from existing works are not clear and are not easy to leverage for cross-dataset metric comparisons.

4. In the future, include diverse summarization domains such as dialogue (Tang et al., 2021; Fabbri et al., 2021a) and email summarization (Zhang et al., 2021), which could potentially have different error types, for a more comprehensive comparison and domain-invariant design of factuality metrics.

We hope that our analysis can shed light on what comparisons practitioners should focus on, how to understand the pros and cons of different metrics, and where metrics should go next.

2 Benchmark

2.1 Benchmark Standardization

Current factuality metrics are evaluated without considering the types of summarization models used to generate the annotated summaries. In these annotated datasets, a large proportion of summaries are generated by older models. Summaries generated by an obsolete model such as a pointer-generator network (See et al., 2017) may contain obvious errors that recent models do not make. **We hypothesize that current factuality systems primarily make progress in identifying factuality inconsistencies from summaries generated by out-of-date summarization models.** If this hypothesis is correct, comparing factuality systems on annotated datasets that contain relatively poor summaries gives us less useful information.

Summarization datasets splits We introduce a new benchmark **AGGREFACT** built on top of Laban et al. (2022). The benchmark **Aggre**gates nine publicly available datasets D that consist of human evaluations of **Fact**ual consistency on model

		AGGREFACT -CNN	AGGREFACT -XSUM
OLD	val	2297	500
	test	2166	430
XFORMER	val	275	500
	test	375	423
SOTA	val	459	777
	test	559	558

Table 1: Statistics of AGGREFACT-CNN and AGGREFACT-XSUM. Details of individual annotated datasets can be found in Appendix Table 5 and 6.

generated summaries. We focus particularly on incorporating recent datasets annotated on top of state-of-the-art pre-trained Transformer models.

All datasets contain summaries generated from articles in CNN/DM and XSum. Given the unique characteristics of CNN/DM and XSum, our proposed benchmark includes two subsets, AGGREFACT-CNN and AGGREFACT-XSUM, that evaluate the performance of factuality metrics on these two datasets separately (Table 1; see also Table 5 and 6 in the Appendix). This can provide more fine-grained and rigorous analysis of the metric performance.

Our benchmark provides factual consistency evaluation via a binary classification task. The binary factual consistency labels for the summaries are determined by human evaluations on the annotated datasets (see details in Section 2.2).

Summarization model splits To validate our hypothesis and make a careful comparison of factuality metrics, we further divide models that were used to generated summaries in the benchmark into three distinct categories: $C = \{$ SOTA, XFORMER, OLD $\}$, as seen in Table 1. SOTA represents state-of-the-art summarization models, including BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020) and T5 (Raffel et al., 2020). XFORMER is a collection of early Transformer-based summarization models. Typical models that fit into this category include BERTSum (Liu and Lapata, 2019), and GPT-2 (Radford et al., 2019). The remaining models, such as Pointer-Generator (See et al., 2017) and BottomUp (Gehrmann et al., 2018), are instances of OLD. A full description of the models in each category is found in Appendix B.

2.2 Benchmark Datasets

In this section, we discuss all datasets that we include in our benchmark. A meta summary of the datasets is shown in Appendix Table 7.

The SUMMAC benchmark (Laban et al., 2022) includes six annotated datasets for factual consistency evaluation. We directly include XSumFaith (Maynez et al., 2020), FactCC (Kryscinski et al., 2020), SummEval (Fabbri et al., 2021b), and FRANK (Pagnoni et al., 2021) from SUMMAC in our benchmark. We do not include the CoGenSumm (Falke et al., 2019) dataset as the original task is ranking pairs of generated summaries instead of detecting factually consistent summaries, and pairs of summaries can be both factually consistent or inconsistent. We modify the Polytope (Huang et al., 2020) dataset in SUMMAC where we view summaries annotated with *addition*, *omission* or *duplication* errors as factually consistent since these three error types are not related to factual consistency. We use the validation and test splits from SUMMAC for the above mentioned datasets.

In addition to modifying SUMMAC, we further include four annotated datasets. For Wang'20 (Wang et al., 2020), CLIFF (Cao and Wang, 2021) and Goyal'21 (Goyal and Durrett, 2021), we create data splits based on the parity of indices, following SUMMAC. For Cao'22 (Cao et al., 2022), we use the existing splits from the original work.

Deduplication and label disagreement correction Some examples may be labeled for errors in multiple datasets. We removed all duplicates so that each instance appears only once in our benchmark. During this deduplication process, we detected 100 instances of the same summaries that are annotated in different datasets with *different* factual consistency labels. 98 of them are between FRANK and XSumFaith, and 2 of them are between FRANK and SummEval. The authors of this work manually corrected the labels for these examples based on our judgment.

2.3 Benchmark Evaluation Metrics

We use balanced accuracy metric to evaluate the performance of factuality metrics due to the imbalance of factually consistent and inconsistent summaries in the benchmark. We refer readers to Laban et al. (2022) for further justification of balanced accuracy as the evaluation metric. In each dataset, a factuality metric selects a threshold for SOTA, XFORMER and OLD, respectively, based on the performance on the corresponding validation set. The chosen thresholds convert raw scores from metrics into binary labels for balanced accuracy evaluation. We provide a weighted average of performance across all datasets in the benchmark (see Table 2).

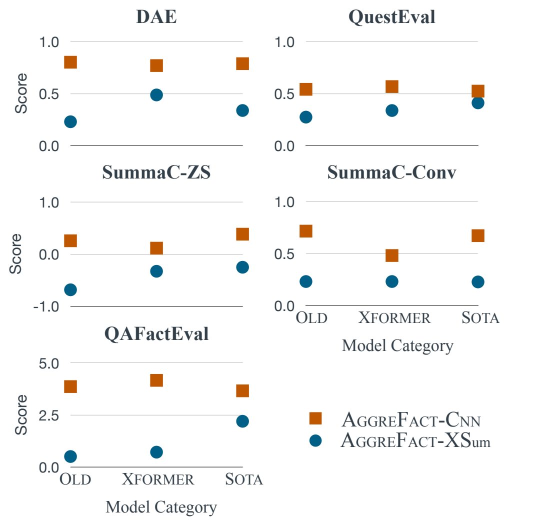

Figure 1: Average threshold values on AGGREFACT-CNN and AGGREFACT-XSUM.

3 Comparison of Factuality Metrics

The first question we approach is **how factuality metrics perform across different datasets**. We re-evaluate several SOTA factual consistency metrics on our benchmark, namely **DAE** (Goyal and Durrett, 2020), **QuestEval** (Scialom et al., 2021), **SummaC-ZS**, **SummaC-Conv** (Laban et al., 2022) and **QAFactEval** (Fabbri et al., 2021c).[2] The full description of these metrics is in Appendix C.

Unifying these metrics We consider each metric as a function $f(d, s) \rightarrow y$, mapping each (document, summary) pair to a score $y \in \mathbb{R}$. For each method, we convert it into a binary classifier $f'(d, s) \rightarrow \{0, 1\}$ by picking a threshold t such that we predict 1 if $f(d, s) > t$ and 0 otherwise.

All thresholds are set separately for each metric. We consider two ways of setting the threshold for a metric: **threshold-per-dataset** and **single-threshold**. The first setting has thresholds $\{t^m_{d,c}\}$ within each metric for every dataset we consider, where d, c and m are any dataset in D, any model category from C, and any factuality metric, respectively. This allows one to choose the right metric

[2] We do not consider other common metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), or BERTScore (Zhang* et al., 2020) as prior work has shown that they do not correlate as well with factual consistency (Fabbri et al., 2021c).

	AggreFact-Cnn			AggreFact-XSum		
	Sota	Xform	Old	Sota	Xform	Old
Baseline	50.0	50.0	50.0	50.0	50.0	50.0
DAE*	59.4	67.9	69.7	73.1	-	-
QuestEval	63.7	64.3	65.2	61.6	60.1	59.7
SummaC-ZS	63.3	76.5	76.3	56.1	51.4	53.3
SummaC-Cv	70.3	69.8	78.9	67.0	64.6	67.5
QAFactEval	61.6	69.1	80.3	65.9	59.6	60.5

Table 2: Weighted evaluation (balanced accuracy) on AggreFact-Cnn and AggreFact-XSum across factuality metrics (threshold-per-dataset setting). Note that a baseline that simply predict all examples as factually (in)consistent can reach a balanced accuracy of 50%. Since DAE was trained on the human-annotated XSumFaith data (Goyal and Durrett, 2021) that includes summaries generated from Xformer and Old, we exclude these summaries for a fair comparison.

	AggreFact-Cnn-Sota	AggreFact-XSum-Sota
DAE	65.4 ± 4.4	$\mathbf{70.2 \pm 2.3}$
QuestEval	$\mathbf{70.2 \pm 3.2}$	59.5 ± 2.7
SummaC-ZS	64.0 ± 3.8	56.4 ± 1.2
SummaC-Conv	61.0 ± 3.9	65.0 ± 2.2
QAFactEval	67.8 ± 4.1	63.9 ± 2.4

Table 3: Balanced binary accuracy using a single threshold on the SOTA subset (single-threshold setting). We show 95% confidence intervals. Highest performance is highlighted in bold.

for the task at hand. The **single-threshold** setting defines one threshold $\{t^c\}$ per metric.

Threshold Analysis We analyze scores from factuality metrics using chosen thresholds $\{t^m_{d,c}\}$ from the validation sets. Specifically, for each factuality metric, we average the values of thresholds for each of Sota, Xformer and Old across all datasets (Figure 1). For all facuality metrics, the average threshold values for AggreFact-Cnn is greater than those for AggreFact-XSum. The discrepancy of threshold values shows that evaluating on both datasets with a single model is a difficult balancing act and may lead to poor results.

We hypothesize that the higher scores from factuality metrics on CNN/DM are related to the extractiveness of the summaries. XSum summaries are more abstractive and tend to contain a larger number of errors, making it harder for the metrics to verify the consistency of summaries with respect to the source text and resulting in lower scores in general. For CNN/DM, smaller deviations from the source may indicate non-factuality.

A weighted average of performance in terms of balanced accuracy for AggreFact-Cnn and AggreFact-XSum is shown in Table 2.[3] We note that for AggreFact-Cnn, factuality metrics achieve the best performance in evaluating the summaries generated from models in Old, with the most recently-introduced metric QAFactEval achieving the highest accuracy of 81.0%. Those summaries contain obvious and obsolete errors that can be more easily detected compared to errors in summaries from more recent models. From Table 1, the majority of annotated summaries are generated by models from Old, so overall performance across datasets will weight these more heavily. However, **there is a significant performance drop when instead evaluating the CNN/DM summaries generated by models from Xformer or Sota.** Approximately a 10% balanced accuracy decrease on average occurs from Old to Sota. Since we mainly use Sota models for text summarization, evaluating the performance of factuality metrics on entire datasets biased towards older models gives us limited information of how these factuality metrics perform on the SOTA-model generated summaries.

In AggreFact-XSum, we do not observe a decrease from Old to Xformer and Sota. Unlike in AggreFact-Cnn, we do not have summaries from a rich set of summarization models from Old and Xformer. As shown in Table 6, only XsumFaith contains less recent model outputs. Since the evaluation already focuses on Sota, there is less of a need for a change in standard empirical practice in this domain.

To encourage future work to compare performance of factuality metric on summaries generated by Sota, we provide a separate benchmark which consists of two subsets AggreFact-CNN-SOTA and AggreFact-XSum-SOTA that only consider summaries generated by Sota models. The validation/test data of AggreFact-CNN-SOTA and AggreFact-XSum-SOTA consists of all validation/test Sota data from AggreFact-Cnn and AggreFact-XSum. This allows the comparisons of factuality metrics using only one threshold.

We show metric comparisons on the SOTA subset in Table 3. Notice that the ranking of factuality metric here (single-threshold setting) is slightly different from the ranking in Table 2 (threshold-

[3]Dataset-wise comparison between factuality metrics is shown in Appendix Table 8.

per-dataset setting). In AggreFact-CNN-SOTA, QuestEval achieves the best performance with no significant difference with most of our evaluated factuality metrics, and DAE performs significantly better on AggreFact-XSum-SOTA. Thus while SummaC-Conv and QAFactEval were in turn proposed as improvements to SOTA on the SummaC benchmark, we find that **metrics which claim improved performance on SummaC do not achieve superior performance when evaluated on Sota summaries.**

4 Finer-grained Error Analysis

Having established differences among factuality metrics across underlying summarization models, we now explore differences in metrics according to factuality error types. To do this, we need a way to unify errors across datasets in our benchmark and map them into a shared taxonomy.

4.1 A Taxonomy of Error Types

We surveyed existing error type taxonomies in prior work and unified the types of factual errors among them into a hierarchical taxonomy in Figure 2. Arrows relate more specific error types to more general "parent" errors. The prior works that make use of each error type can be found in Appendix D. As shown in the figure, most error types related to factual consistency fall under the subset *{intrinsic, extrinsic}* × *{noun phrase, predicate}* if we consider the coarsest level of the hierarchy. We discard discourse errors as these are uncommon and not available in most of our datasets. Therefore, we unify unique error type taxonomies from all four datasets we consider here into this error type subset (shown in the gray box in Figure 2). Descriptions and examples for these error types are in Table 9. Further, we introduce two additional error categories *{intrinsic-entire sent., extrinsic-entire sent.}* if the entire summaries are annotated as having hallucinations.

We are able to map four of the datasets in AggreFact that contain fine-grained annotations to our unified taxonomy. For all four datasets, if there are multiple annotators, we assign an error type to a summary if the error is annotated by more than one annotator, and we allow one summary to have multiple error types. We call the annotated subset related to CNN/DM and XSum as AggreFact-Cnn-Unified and AggreFact-XSum-Unified, respectively.

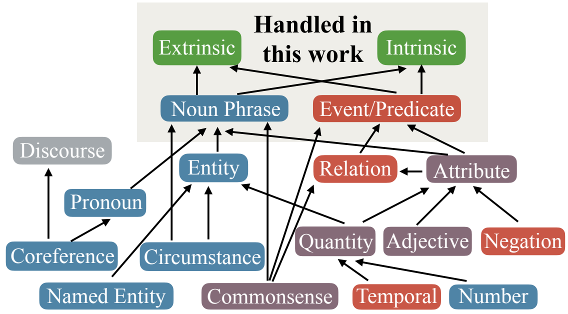

Figure 2: Taxonomy of factual consistency errors. We use unique colors to represent entity- and predicate-related errors, as well as the mix of two. See Appendix D for citations of papers that use each error type.

4.2 Error Mapping

XSumFaith XSumFaith consists of 500 summaries each from human reference, two models in Old, and two models in XFormer. All summaries are annotated with intrinsic and extrinsic errors, but no finer categories are distinguished. To perform error type mapping, we detect predicates in a summary and assign each hallucinated text span intrinsic- or extrinsic-predicate error if it contains a predicate. We map the remaining hallucinated spans to intrinsic- or extrinsic-noun phrase error.

FRANK The CNN/DM subset of FRANK consists of three models in Old, and one model each in both XFormer and Sota. The XSum portion of FRANK has two models each in Old and XFormer. Each model contains 250 summaries in the dataset. We mapped Entity error and Out of Article error to extrinsic-noun phrase error; Predicate error and Grammatical error to extrinsic-predicate error; Circumstance error and Coreference error to intrinsic-noun phrase error; and other errors to intrinsic-predicate error.

Goyal'21 Authors of the original dataset manually identified all hallucinated text spans for each summary and classified hallucination types into {intrinsic, extrinsic} × {entity, event, noun phrase, others}. The dataset consists of summaries for both CNN/DM and XSum. For the CNN/DM susbset, the authors directly annotated 50 summaries from FactCC, where summaries were generated by Old models. The XSum subset consists of summaries from Sota models. We map entity-related and noun phrase-related errors to noun phrase errors, event errors to predicate errors and others to entire sentence errors.

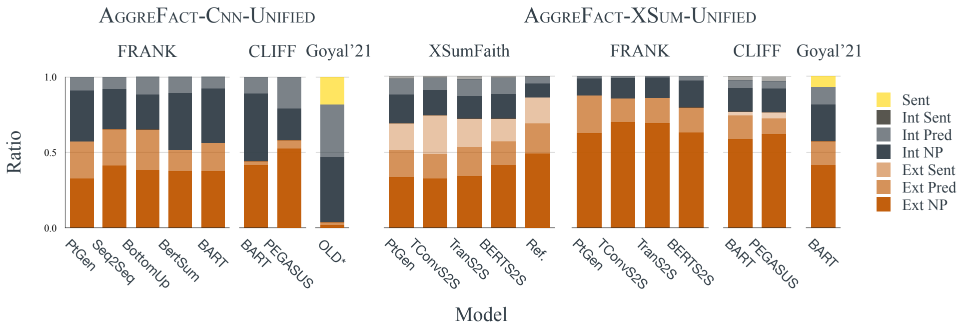

Figure 3: Error types of summaries from AGGREFACT-CNN-UNIFIED and AGGREFACT-XSUM-UNIFIED. Ref. is annotated reference summary from XSumFaith. Since Goyal'21 in AGGREFACT-CNN-UNIFIED annotated summaries from FactCC, we use OLD* to denote summaries generated from OLD models.

CLIFF This dataset consists of 150 summaries each for both CNN/DM and XSum from two models in SOTA. We use the same approach for error mapping as we do for XSumFaith by only considering words labeled as extrinsic or intrinsic errors.

We evaluate the accuracy of our error type mapping via manual inspection. Specifically, the authors of this work inspect 30 factually inconsistent examples each for XSumFaith, FRANK and CLIFF. Those examples cover summaries generated by all models used in the datasets. Results of the manual inspection show that the accuracy of our error type mapping is over 90%.

A common discrepancy noticed by annotators was that in several cases the examples were originally annotated as intrinsic/extrinsic but we believe those errors are extrinsic/intrinsic. These cases, however, are not a result of any error in our mapping, but instead disagreement or error in the original annotation itself. We found that our error mapping for FRANK is not as accurate as for the remaining three datasets. For example, we found that the entity error (EntE) can be either intrinsic or extrinsic even though the FRANK authors have defined "out of article" error, which could be noun phrase or predicate errors as well. Since the definitions of error types in Goyal'21 closely resemble our mapping and there are 150 examples in total, we correct any errors in the mapping on this dataset. Corrections mostly happens for the event-related error defined in Goyal'21 in that event-related error can be either noun phrase-related or predicate-related.

4.3 Distribution Shift of Error Types

Next, we explore how the number of errors in specific groups of models from SOTA, XFORMER, and OLD has changed with the progress in the field.

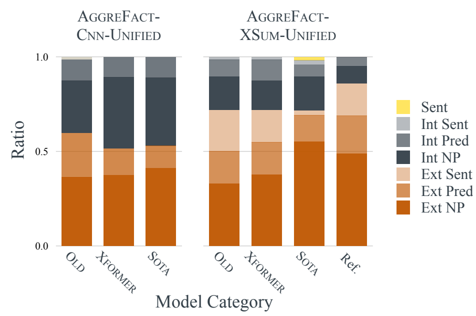

Figure 4: Distribution shift of error types on AGGREFACT-CNN-UNIFIED and AGGREFACT-XSUM-UNIFIED. Ref. is human reference from XSumFaith.

Specifically, for each of the FRANK, XSumFaith, Goyal'21, and CLIFF datasets, we calculate the ratio of error types from factually inconsistent summaries generated by each model. We then study any distribution shift of error types in AGGREFACT-CNN-UNIFIED and AGGREFACT-XSUM-UNIFIED under SOTA, XFORMER, and OLD.

Summaries generated by the same models consist of different error distributions over different datasets. As shown in AGGREFACT-XSUM-UNIFIED (Figure 3), BART summaries are annotated by both Goyal'21 and CLIFF. However, it is interesting that BART summaries were annotated as making more intrinsic-noun phrase and intrinsic-predicate errors in Goyal'21 but more extrinsic-noun phrase errors in CLIFF. Similar observations can be found in AGGREFACT-CNN-UNIFIED, where BART summaries have a higher proportion of extrinsic-predicate error in FRANK and more intrinsic-noun phrase error in CLIFF.

In addition, although XSumFaith and FRANK annotate the same set of model generated summaries in AGGREFACT-XSUM-UNIFIED, the distribution of error types looks dramatically differ-

	AGGREFACT-CNN-ERROR				AGGREFACT-XSUM-ERROR					
	Intrinsic		Extrinsic		Intrinsic			Extrinsic		
	NP (183)	Pred. (60)	NP (220)	Pred. (129)	NP (196)	Pred. (113)	Sent (17)	NP (434)	Pred. (181)	Sent (197)
DAE*	59.6	53.3	67.7	62.8	-	-	-	-	-	-
QuestEval	62.8	50.0	72.3	68.2	33.2	44.2	64.7	40.6	50.3	69.0
SummacZS	66.1	**71.7**	**81.8**	72.1	50.0	57.5	76.5	48.6	47.5	36.0
SummacConv	62.8	**65.0**	76.4	59.7	54.1	62.8	29.4	**64.5**	60.8	70.6
QAFactEval	56.3	51.7	**79.1**	63.6	**66.8**	**75.2**	**88.2**	55.1	**70.2**	**79.2**

Table 4: Recall of identified hallucinated summaries that contain certain error types across datasets (XSumFaith, FRANK, Goyal'21 and CLIFF) and factuality metrics. Binary labels are directly obtained from AGGREFACT-CNN and AGGREFACT-XSUM. Numbers of summaries that have certain error types are shown in the parentheses. We obtain 95% confidence intervals and numbers in **bold** indicates that models have significantly higher recall of identifing certain error types compared to the rest of of the metrics. Since DAE is trained with human annotated data from XSumFaith, we remove DAE for a fair comparison in XSum error types.

ent. The main discrepancy lies in the proportion of extrinsic-noun phrase and intrinsic-predicate errors. There are two possible reasons for such discrepancy. First, FRANK does not have "entire sent." errors as it only contains sentence-level annotations. Second, and more important, it is not easy to map error types from FRANK directly to our unified error types in spite of our validation. For example, the "out of article error" in FRANK is defined as an error where some statements in the summary do not show up in the source text. We found this error can be mapped to either an extrinsic-noun phrase error or extrinsic-predicate error. These observations indicate that **previous work disagrees about where the individual error class boundaries are, even when aligned with our taxonomy**.

A combined meta-analysis shows shifts in error distributions. Figure 3 show that in each annotated dataset the error type distribution may vary among models from the same category. For example, summaries from BART contain a higher ratio of intrinsic-noun phrase errors than summaries from PEGASUS in AGGREFACT-CNN-UNIFIED. We now combine all datasets together from AGGREFACT-CNN-UNIFIED and AGGREFACT-XSUM-UNIFIED and show the unified error distributions over three model categories.[4] As shown in Figure 4, models make approximately 50% extrinsic errors in CNN/DM, with a slightly decrease from OLD to more recent models. For XSum, the proportion of extrinsic errors remains unchanged and are at 70%. SOTA models generate a higher proportion of intrinsic errors for CNN/DM and a higher proportion of extrinsic errors for XSum. This observation aligns with our intuition as CNN/DM is more extractive, and XSum is highly abstrative and contains large amount of hallucinated human reference summaries. Within extrinsic errors in XSum, more recent models generate less completely wrong summaries.

[4]For AGGREFACT-XSUM-UNIFIED, since XSumFaith and FRANK annotated the same set of summaries, we only use the annotation results from XSumFaith since our error mapping is more accurate on the span-level annotations.

4.4 Error Type Detection by metrics

In this section, we analyze how factuality metrics perform on summaries that contain certain error types. Specifically, we collect subsets of examples from four annotated datasets and group them into AGGREFACT-CNN-ERROR and AGGREFACT-XSUM-ERROR.[5] Every subset contains summaries that include one error type defined in Section 4.1. Each factuality metric assigns a binary label to an instance obtained directly from AGGREFACT-CNN and AGGREFACT-XSUM. Note that each subset only consists of test set examples from our benchmark since examples from the validation set were used to choose the optimal thresholds (Section 3). Since there are limited annotations for each model category after only considering examples from the test set of the benchmark, we decide not to split data by model categories in this part of the analysis. We calculate the recall of identifying error types from those subsets and show the results in Table 4. Note that the performance of DAE is excluded for AGGREFACT-XSUM-ERROR since DAE is trained with human annotations from XSumFaith.

Summaries from AGGREFACT-CNN-ERROR and AGGREFACT-XSUM-ERROR primarily come

[5]We exclude FRANK for this analysis for the same reason as in Section 4.3.

from non-SOTA models (89.6% and 92.1%, respectively). On AGGREFACT-CNN-ERROR, where 79.0% of summaries were generated from OLD, there are more extrinsic errors (349) than intrinsic errors (243). This follows our above analysis as errors from more than 50% of summaries generated by less recent models are extrinsic (Figure 4).

Across AGGREFACT-CNN-ERROR and AGGREFACT-XSUM-ERROR, we found that SummaC-Conv and QAFactEval achieve higher recall for most error types. This indicates that **more recent factuality metrics are better at capturing obsolete errors generated from less recent models.** This observation aligns with our finding in Table 2 (column EARLY-TRANS and OLD) in general. Interestingly, we find that **summarization datasets (CNN/DM and XSum) have a non-negligible effect on the metrics' capabilities of detecting certain error types, even in the cases of out-of-date errors.** For example, the recall of identifying extrinsic-noun phrase error drops 10-30% across all factuality metrics when evaluated on AGGREFACT-XSUM-ERROR, and multiple models perform worse in general on identifying errors from AGGREFACT-XSUM-ERROR. Another observation is that although DAE is trained using annotations from XSumFaith, it does not identify errors as well in AGGREFACT-CNN-ERROR. These findings indicate that **summarization models make fundamentally different errors for each error type, and current factuality metrics cannot be uniformly good at identifying certain error types across datasets.** We believe this conclusion still holds when evaluating metrics on summaries generated from SOTA models since they generate less obvious errors.

5 Recommendations

Evaluate factuality models on modern systems We have seen that SOTA yields significantly different results than XFORMER or OLD. Because of the prevalence of these systems, we believe that any new work should prefer evaluating on these SOTA datasets. Particularly for factuality methods that use pre-trained models, evaluating on pre-trained summarizers is needed to see if these metrics are improving from the current state-of-the-art or merely patching errors in outdated systems that have already been fixed by other advances.

Choose the right metric for the job We note that there is no one clear winner among the metrics evaluated here (Section 3). Depending on the downstream application, different methods may be more or less appropriate, as our analysis shows. An ensembling of different methods or a metric that combines the merits of existing metrics may bring additional performance boost. Moreover, none of current factuality metrics can identify certain error types across datasets equally well. As QG/QA and NLI models get better, we expect all of these methods to improve further.

Use more consistent error types With our taxonomy, we have mapped error types annotated in previous work. It is relatively easier and more accurate to map errors from XSumFaith, Goyal'21, and CLIFF to our unified error types as they have annotation granularity finer than sentence-level. We encourage future work to follow this taxonomy where possible and leverage definitions in prior work to improve the potential to make *cross-dataset* comparisons. To evaluate which error type a factuality metric is good at identifying, we encourage future work to annotate and evaluate specifically on SOTA model generated summaries.

Annotate and evaluate on non-news datasets Most of current annotated datasets are within the news domain and factuality metrics are evaluated on news summaries accordingly. As there is a rising interest in other domains such as dialogue summarization (Tang et al., 2021; Fabbri et al., 2021a) and email summarization (Zhang et al., 2021), future work could annotate and analyze errors made by SOTA models there. We encourage future work to develop factuality metrics that have superior performance over cross-domain evaluation.

6 Conclusion

In this work, we analyzed several factuality metrics across a large meta-benchmark assembled from existing datasets. We find that state-of-the-art summarization models still present challenges for detecting factual errors, and the performance of error detectors is often overestimated due to the reliance on older datasets. Furthermore, we unify existing datasets into a common taxonomy and use this to highlight differences between datasets and summarization models, as well as the complexity of unifying concepts in this problem space.

References

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

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

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.

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

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. 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, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021a. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. 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 6866–6880, Online. Association for Computational Linguistics.

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

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021c. Qafacteval: Improved qa-based factual consistency evaluation for summarization.

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.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. 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.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141, Melbourne, Australia. Association for Computational Linguistics.

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.

Yichen Jiang and Mohit Bansal. 2018. Closed-book training to improve summarization encoder memory. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4067–4077, Brussels, Belgium. Association for Computational Linguistics.

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

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. 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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. 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.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press.

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

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021b. Improving factual consistency of abstractive summarization via question answering. 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 6881–6894, 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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ramakanth Pasunuri and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. 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 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*.

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

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.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning.

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

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.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Zhiyuan Zeng, Jiaze Chen, Weiran Xu, and Lei Li. 2021. Gradient-based adversarial factual consistency evaluation for abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4102–4108, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. EmailSum: Abstractive email thread summarization. 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 6895–6909, Online. Association for Computational Linguistics.

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

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

A Limitations

There are a few limitations of our work. First, we focus on evaluating state-of-the-art factuality metrics on English newswire datasets. This setting restricts us to English-language data, a formal style of text, and topics consisting of what is discussed in US and UK-centric news sources. Moreover, other summarization domains such as dialogue summarization have different common error types such as *wrong reference error* (Tang et al., 2021), which are not fully evaluated under current metrics. As settings like this are studied in future work, we believe that the kinds of analysis we do here can be extended to these settings as well.

Second, since our work is built on top of previous work, some analysis such as the error type mapping is limited by the quality and annotation agreement from previous work. We chose not to undertake large-scale reannotation to avoid causing confusion in the literature with multiple versions of datasets reflecting divergent annotator opinions. In spite of these limitations, we believe that our re-evaluation of these metrics and the analysis of error types under newswire data can bring insights for future works in choosing, designing and evaluating factuality metrics.

B Model Categories

In this section, we briefly describe the summarization models we use in this paper.

For SOTA, we include Transformer-based pre-trained models like BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and PEGASUS (Zhang et al., 2020). They are pre-trained on massive text corpus and further fine-tuned on summarization datasets.

For XFORMER, we use BERTSumExt and BERTSumAbs from Liu and Lapata (2019), GPT-2 (Radford et al., 2019), TransS2S (Vaswani et al., 2017), and BERTS2S (Devlin et al., 2019).

For OLD, we include models FastAbsRl (Chen and Bansal, 2018), TConvS2S (Narayan et al., 2018), BottomUp (Gehrmann et al., 2018), PGNet (See et al., 2017), NeuSUM (Zhou et al., 2018), BanditSum (Dong et al., 2018), SummaRuNNer (Nallapati et al., 2017), TextRank (Mihalcea and Tarau, 2004), CBDec (Jiang and Bansal, 2018), RNES (Wu and Hu, 2018), ROUGESal (Pasunuru and Bansal, 2018), ImproveAbs (Kryściński et al., 2018), MultiTask (Guo et al., 2018), and UnifiedExtAbs (Hsu et al., 2018).

C Factuality Metrics

We show the descriptions of consistency metrics we considered in our benchmark.

DAE Goyal and Durrett (2020) propose an arc entailment approach that evaluates the factuality $F_a(a, x) = P(\text{entailment} \mid a, x)$ of each dependency arc $a \in \text{Arc}(s)$ of the generated summary s independently with respect to the input article x. It then uses their aggregation $\frac{1}{|\text{Arc}(s)|}\sum_{a\in\text{Arc}(s)} F_a(a, x)$ as the overall score. We use the default model and hyperparameters provided by the authors,[6] described in Goyal and Durrett (2021), which is trained on data from XSum-Faith, which we account for later in our comparisons.

QuestEval Scialom et al. (2021) propose a QA-based metric that aggregates answer overlap scores from selected spans r and questions $q_i \in Q_G(x)$ that derived from the input article x and answered $Q_A(s, q_i)$ using the summary s (recall-based); and those derived from the summary $q_i \in Q_G(s)$ and answered $Q_A(x, q_i)$ using the input article x (precision-based). Q_G and Q_A denote question generation and question answering components, respectively. We use the implementation provided by the authors[7] and apply the unweighted version of the metric as in Laban et al. (2022).

SummaC-ZS Laban et al. (2022) is a zero-shot entailment metric that computes a sentence-level entailment score $F(s_i, x_j)$ between each summary sentence s_i and input sentence x_j using an NLI model F. It first find the maximum entailment score $\text{score}(s_i) = \max_j F(s_i, x_j)$ for each summary sentence s_i, and averaging over all summary sentences for the final score $\frac{1}{|s|}\sum_i \text{score}(s_i)$. We use the default model and hyperparameters provided by the authors, which may return a negative score.

SummaC-Conv Laban et al. (2022) extends SummaC-ZS by replacing the max operation with a binning of the entailment scores between each summary sentence s_i and all input sentences x_j to create a histogram $\text{hist}(s_i, x)$. The histogram is then passed through a learned 1-D convolution layer Conv to produce the summary sentence score $\text{score}(s_i) = \text{Conv}(\text{hist}(s_i, x))$. Parameters for the convolution layer are learned on synthetic data from FactCC (Kryscinski et al., 2020).

QAFactEval Fabbri et al. (2021c) is a QA-based metric analogous to the precision-based component of QuestEval and includes optimized question answering, generation, and answer-overlap components. We do not make use of the variation of

[6] https://github.com/tagoyal/factuality-datasets

[7] https://github.com/ThomasScialom/QuestEval

QAFactEval which combines QA and entailment-based scores into a single metric.

D Surveyed Error Types

Here are our surveyed error types that are related to factual inconsistency.

Negation Error (Zhang et al., 2020; Kryscinski et al., 2020; Huang et al., 2020; Zeng et al., 2021)

Adjective Error (Zhang et al., 2020)

Coreference Error (Zhang et al., 2020; Kryscinski et al., 2020; Pagnoni et al., 2021; Nan et al., 2021b)

Number error (Kryscinski et al., 2020; Nan et al., 2021b; Chen et al., 2021; Cao et al., 2020)

Entity error (Kryscinski et al., 2020; Pagnoni et al., 2021; Zeng et al., 2021; Wang et al., 2020; Nan et al., 2021b,a; Chen et al., 2021; Cao et al., 2020)

Attribute error (Pagnoni et al., 2021; Huang et al., 2020)

Pronoun error (Kryscinski et al., 2020; Zeng et al., 2021; Cao et al., 2020)

Commonsense error (Kryscinski et al., 2020)

Temporal error (Kryscinski et al., 2020; Cao et al., 2020)

Predicate error (Pagnoni et al., 2021)

Discourse link Error (Pagnoni et al., 2021)

Relation error (Nan et al., 2021a,b)

Quantity error (Zhao et al., 2020)

Event error (Goyal and Durrett, 2021),

Noun phrase error (Wang et al., 2020; Goyal and Durrett, 2021),

Circumstance error (Pagnoni et al., 2021)

		Polytope	FactCC	SummEval	FRANK	Wang'20	CLIFF	Goyal'21	Total
OLD	val	450	931	550	223	118	-	25	2297
	test	450	503	548	523	117	-	25	2166
XFORMER	val	150	-	50	75	-	-	-	275
	test	150	-	50	175	-	-	-	375
SOTA	val	34	-	200	75	-	150	-	459
	test	34	-	200	175	-	150	-	559

Table 5: Statistics of AGGREFACT-CNN. Each dataset is stratified into three categories OLD, XFORMER, and SOTA.

		XsumFaith	Wang'20	CLIFF	Goyal'21	Cao'22	Total
OLD	val	500	-	-	-	-	500
	test	430	-	-	-	-	430
XFORMER	val	500	-	-	-	-	500
	test	423	-	-	-	-	423
SOTA	val	-	120	150	50	457	777
	test	-	119	150	50	239	558

Table 6: Statistics of AGGREFACT-XSUM.

Dataset	**Annotators**	**Kappa**	**Gran**	**Annotation Scheme**
FactCC (Kryscinski et al., 2020)	2 authors	-	summ	binary consistency label (consistent/inconsistent)
Wang'20 (Wang et al., 2020)	3 crowd-sourced annotators	0.34/0.51	sent	binary consistency label (consistent/inconsistent)
SummEval (Fabbri et al., 2021b)	5 crowd-sourced annotators and 3 authors	0.70	summ	5-point Likert scale
Polytope (Huang et al., 2020)	3 trained annotators	-	span	{addition, ommision, inaccuracy intrinsic, inaccuracy extrinsic, positive-negative aspect}
Cao'22 (Cao et al., 2022)	2 authors and 3 graduate students	0.81	entity	{Non-hallucinated, Non-factual Hallucination, Intrinsic Hallucination, Factual Hallucination}
XSumFaith (Maynez et al., 2020)	3 trained annotators	0.80	span	{intrinsic, extrinsic}
FRANK (Pagnoni et al., 2021)	3 crowd-sourced annotators	0.53	sent	{RelE, EntE, CircE, OutE, GramE, LinkE, CorefE, OtherE, NoE}
Goyal'21 (Goyal and Durrett, 2021)	2 authors	-	span	{intrinsic, extrinsic} $\times$ {entity, event, noun phrase, others}
CLIFF (Cao and Wang, 2021)	2 experts	0.35/0.45	word	{intrinsic, extrinsic, world knowledge, correct}

Table 7: Metadata of nine datasets in the benchmark. We report the source of annotators, inter-annotator aggrement, annotation granularity, and annotation scheme for each dataset. Wang'20 and CLIFF reported kappa scores for XSum/CNNDM seperately.

	Dataset	Category	Count	Factuality Metric				
				DAE	QuestEval	SummaC-ZS	SummaC-Conv	QAFactEval
CNN/DM	FactCC	OLD	503	0.704	0.655	0.835	0.891	0.843
	Wang'20	OLD	117	0.586	0.552	0.655	0.672	0.754
	SummEval	OLD	548	0.661	0.649	0.773	0.801	0.815
		XFORMER	50	0.760	0.680	0.620	0.580	0.740
		SOTA	200	0.452	0.649	0.622	0.827	0.652
	Polytope	OLD	450	0.779	0.687	0.802	0.791	0.824
		XFORMER	150	0.774	0.733	0.970	0.811	0.726
		SOTA	34	0.294	0.176	0.971	0.735	0.324
	FRANK	OLD	523	0.704	0.669	0.692	0.728	0.773
		XFORMER	175	0.574	0.556	0.631	0.634	0.646
		SOTA	175	0.699	0.626	0.570	0.601	0.547
	Goyal'21	OLD	25	0.188	0.146	0.375	0.354	0.271
	CLIFF	SOTA	150	0.730	0.740	0.646	0.649	0.716
XSum	Wang'20	SOTA	119	0.756	0.560	0.698	0.721	0.756
	Cao'22	SOTA	239	0.723	0.601	0.490	0.668	0.613
	XSumFaith	OLD	430	-	0.597	0.533	0.675	0.605
		XFORMER	423	-	0.601	0.514	0.646	0.596
	Goyal'21	SOTA	50	0.644	0.814	0.466	0.552	0.754
	CLIFF	SOTA	150	0.754	0.619	0.596	0.668	0.613

Table 8: Dataset-wise comparsion between factuality metrics. Since DAE is trained with human annotated data from XsumFaith, we remove DAE for a fair comparison.

Error Type	Definition	Example of Generated Summaries
Intrinsic-Noun Phrase	A model misrepresents word(s) from the source text that function(s) in a summary as subject, object, or prepositional object.	The world's first subsea power hub which uses a lithium-based drive system to generate electricity is being tested off the west coast of orkney.
Intrinsic-Predicate	A model misrepresents word(s) from the source text that function(s) in a summary as the main content verb or content like adverbs that closely relate to the verb.	A conservative mp has resigned from his constituency as part of an investigation into a # 10.25 m loan to a football club.
Extrinsic-Noun Phrase	A model introduces word(s) not from the source text that function(s) in a summary as subject, object, or prepositional object but cannot be verified from the source.	Shale gas drilling in lancashire has been suspended after a magnitude-7.5 earthquake struck.
Extrinsic-Predicate	A model introduces word(s) not from the source text that function(s) in a summary as the main content verb or content like adverbs that closely relate to the verb, but which cannot be verified from the source.	Folate - also known as folic acid - should be added to flour in the uk, according to a new study.

Table 9: Definition and examples of unified error types. Factually inconsistent spans are highlighted in red.