# Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics

**Artidoro Pagnoni** **Vidhisha Balachandran** **Yulia Tsvetkov**
Language Technologies Institute
Carnegie Mellon University
{apagnoni,vbalacha,ytsvetko}@cs.cmu.edu

## Abstract

Modern summarization models generate highly fluent but often factually unreliable outputs. This motivated a surge of metrics attempting to measure the factuality of automatically generated summaries. Due to the lack of common benchmarks, these metrics cannot be compared. Moreover, all these methods treat factuality as a binary concept and fail to provide deeper insights on the kinds of inconsistencies made by different systems. To address these limitations, we devise a *typology of factual errors* and use it to collect human annotations of generated summaries from state-of-the-art summarization systems for the CNN/DM and XSum datasets. Through these annotations we identify the proportion of different categories of factual errors in various summarization models and benchmark factuality metrics, showing their correlation with human judgement as well as their specific strengths and weaknesses.[1]

## 1 Introduction

Factuality is defined as a measure of "whether eventualities are characterized as corresponding to facts, possibilities, or situations that do not hold in the world" (Sauri, 2008; Saurí and Pustejovsky, 2012). In summarization, this "world" is the article, which is taken as ground-truth, and the output summary must be faithful to the article's facts. Despite advancements in neural abstractive summarization (Narayan et al., 2018; Liu and Lapata, 2019; Lewis et al., 2020), ∼30% of summaries have factual inconsistencies (Cao et al., 2018). With summarization being an integral component of information consumption, this highlights a need for ensuring summarization systems are factually consistent and developing methods for evaluating them.

Common evaluation metrics for summarization based on n-gram overlap – BLEU, ROUGE, and METEOR (Papineni et al., 2002; Lin, 2004; Lavie and Agarwal, 2007) – are insufficient to measure the factual correctness of summaries and fail to correlate with the human judgements of factuality (Falke et al., 2019; Kryscinski et al., 2019). More recent metrics proposed to improve the evaluation of summarization factuality (Kryscinski et al., 2020; Durmus et al., 2020; Wang et al., 2020; Maynez et al., 2020) cannot be compared due to the lack of common benchmarks. More critically, while these approaches differ in the way they model factuality, they all consider factuality as a binary concept, labeling summaries of any length as factual or non-factual. They do not provide any fine-grained understanding of the factual errors made by different systems that could serve as an actionable feedback on a system's limitations.

The binary factuality of a text can be difficult to determine. Falke et al. (2019) show relatively low crowd–expert agreement, indicating the presence of subjectivity in the annotation process. Moreover, not all factual errors are equally important and the number of errors can have a significant impact on the perceived factuality of a text. This suggests that non-factuality should be modeled as a multi-dimensional construct and not a label.

In this work, we propose a linguistically motivated typology of factual errors for fine-grained analysis of factuality in summarization systems (§2). Our typology is theoretically grounded in frame semantics (Fillmore et al., 1976; Palmer et al., 2005) and linguistic discourse theory (Brown and Yule, 1983). It provides several benefits. First, we find that decomposing the concept of factuality in (relatively) well-defined and grounded categories makes the final binary decision more objective leading to near perfect agreement between crowd and expert annotators ($\kappa = 0.86$). Second, this approach provides some measure of the degree of non-factuality both in terms of the quantity and the category of factual violations that appear

[1]Code, data, and online leaderboard will be available at https://github.com/artidoro/frank



*Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829
June 6–11, 2021. ©2021 Association for Computational Linguistics

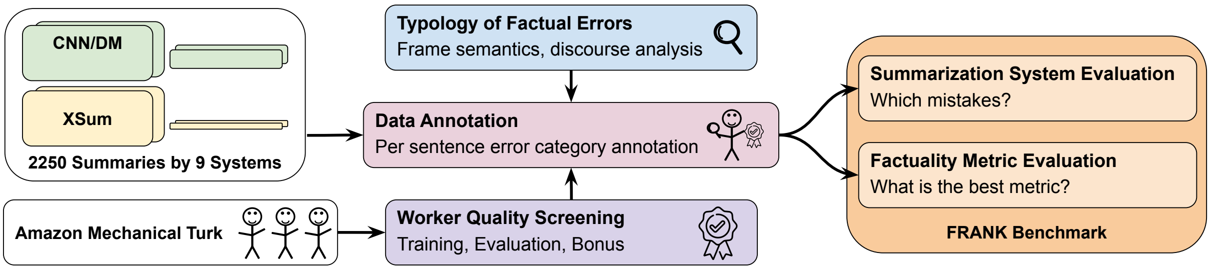


Figure 1: We propose a linguistically grounded typology of factual errors. We select crowd workers to annotate summaries from two datasets according to this typology achieving near perfect agreement with experts. We collect FRANK, the resulting dataset, to benchmark factuality metrics and state-of-art summarization systems.

in the text. This typology also provides us with the means to categorize the types of errors made by summarization systems, helping us gain deeper insights than simply categorizing content as factual or hallucinated.

We define an annotation protocol of factuality based on our typology and collect a dataset of human judgements over a diverse set of model generated summaries on the CNN/DM (Hermann et al., 2015) and XSum (Narayan et al., 2018) datasets (§3). Through this dataset, we aim to both assess the factuality of summarization systems and benchmark recently proposed factuality metrics. In §4 we discuss various state-of-art models and show a detailed analysis of the factual errors they make. Finally, in §5 we evaluate multiple summarization metrics against our benchmark and show their strengths and weaknesses in detecting specific types of factual errors. Figure 1 shows an overview of this work.

## 2 Typology of Factual Errors

Previous studies of factuality in summarization only distinguish factual and hallucinated content (Kryscinski et al., 2019; Maynez et al., 2020) and provide limited insights on the fine-grained types of factual errors. In the simplest case, factual errors appear within a single proposition. However, as summaries include several sentences, discourse markers describe relations across propositions. These cross-sentence links, such as causality or temporal ordering, can introduce inconsistencies with the article. Furthermore, information in the summary should be verifiable given the article. This understanding outlines different levels of linguistic structure where factual mistakes can arise in summaries: at the semantic frame level, at the discourse level, or because the content cannot be verified. Below we define a typology of factual errors further detailing these three levels. This typology is theoretically grounded in frame semantics (Fillmore et al., 1976; Baker et al., 1998; Palmer et al., 2005) and linguistic discourse analysis (Brown and Yule, 1983). Examples for each category are shown in Table 1.

### 2.1 Semantic Frame Errors

A *semantic frame* is a schematic representation of an event, relation, or state, which consists of a predicate and a list of participants, called frame elements (Baker et al., 1998). A semantic frame has both core and non-core frame elements (FE). Core frame elements are essential to the meaning of the frame, while non-core (e.g. location, time) provide additional descriptive information. Our first three categories capture factual errors in each of these components (frame, core and non-core FE) respectively.

**Predicate Error (PredE):** Category *PredE* encompasses errors where the predicate in a summary statement is inconsistent with the source text. More generally, this represents cases where the frame from a summary statement does not align with what is expressed in the source text.

**Entity Error (EntE):** Category *EntE* captures errors where the primary arguments (like entities) of the predicate are wrong or have the wrong attributes, although the relation was expressed in the original text. More generally, these account for cases where the core frame elements in a frame are wrong. This also captures directionality errors where the elements are interchanged (similar to agent-patient swap).

**Circumstance Error (CircE):** In additional to the core arguments, predicates can be further specified using additional information or attributes that describe the circumstance in which the arguments

| | Category | Description | Example |
|---|---|---|---|
| **PredE** | Relation Error | The predicate in the summary statement is inconsistent with the source article. | *The Ebola vaccine **was rejected** by the FDA in 2019.* |
| **EntE** | Entity Error | The primary arguments (or their attributes) of the predicate are wrong. | *The **COVID-19 vaccine** was approved by the FDA in 2019.* |
| **CircE** | Circumstance Error | The additional information (like location or time) specifying the circumstance around a predicate is wrong. | *The first vaccine for Ebola was approved by the FDA in **2014**.* |
| **CorefE** | Coreference Error | A pronoun/reference with wrong or non-existing antecedent. | *The first vaccine for Ebola was approved in 2019. **They** say a vaccine for COVID-19 is unlikely to be ready this year.* |
| **LinkE** | Discourse Link Error | Error in how multiple statements are linked together in the discourse (for example temporal ordering/causal link). | *To produce the vaccine, scientists have to show successful human trials, **then** sequence the DNA of the virus.* |
| **OutE** | Out of Article Error | The statement contains information not present in the source article. | ***China** has already started clinical trials of the COVID-19 vaccine.* |
| **GramE** | Grammatical Error | The grammar of the sentence is so wrong that it becomes meaningless. | *The Ebola vaccine **accepted have already started.*** |

Table 1: Typology of factual errors. Original text for the examples: *The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.*

and predicates interact (e.g. location, time, manner, direction, modality). Category *CircE* captures errors where one or more such attributes (non-core frame elements within a frame) are wrong.

### 2.2 Discourse Errors

The communicative intent of an author is also expressed through relations that hold between parts of the text. Factual errors in summarized text can often extend beyond a single semantic frame introducing erroneous links between discourse segments. Below we outline such categories of errors which are grounded in discourse analysis and rhetorical structure theory (RST) (Brown and Yule, 1983; Mann and Thompson, 1988). RST is an elaborate system for annotating coherence relations in discourse. Some examples of such relations include: "Elaboration", "Background", "Motivation", and "Volitional Cause". Here we depart from semantic frame terminology as its rooting in a single frame does not allow us to represent such errors.

**Coreference Error (CorefE):** Category *CorefE* accounts for errors where pronouns and other types of references to previously mentioned entities either are incorrect or have no clear antecedents, making them ambiguous.

**Discourse Link Error (LinkE):** Category *LinkE* encompasses errors involving a discourse link between different statements. These include errors of incorrect temporal ordering or incorrect discourse links (e.g. RST relations, discourse connectors) between statements.

### 2.3 Content Verifiability Errors

Often statements in a summary cannot be verified against the source text due to difficulty in aligning them to the source. Below we outline two categories of errors for such cases.

**Out of Article Error (OutE):** Since summaries of a document should only contain information that can be deduced from the original text, we include a category for such errors *OutE* (prior work refers to this as extrinsic hallucinations (Maynez et al., 2020)).

**Grammatical Error (GramE):** We use *GramE* to categorize statements that are not well formed. When grammatical mistakes make the meaning of a statement incomprehensible or ambiguous, it cannot be verified against the source and is thus considered trivially wrong. Minor grammatical errors are acceptable.

Finally, for completeness in our annotation exercise, we add two additional categories **Others (OthE)** for factually errors that do not correspond to any of the above categories and **Not an Error (NE)** for statements that do not contain any errors.

## 3 Dataset Creation

Beyond theoretical grounding, we empirically verify our typology through large scale human annotations of five abstractive summarization models on the CNN/DM dataset and four on the XSum dataset. Through our dataset, we aim to have a broad coverage of different types of errors made by neural summarization systems, with human judgements on their fine-grained factuality errors.

**Annotation Data** For the annotation, we include model summaries from CNN/DM and XSum datasets as they present different characteristics. CNN/DM summaries are longer, with three sentences on average, while XSum has only single sentence summaries. Having longer summaries is crucial to identify discourse level errors. On the other hand, XSum summaries are more abstractive and include more factual errors on average (Maynez et al., 2020). For a diverse set of model summaries, we collect publicly available model outputs from different summarization models with differing factuality capabilities. For the CNN/DM dataset, we use model outputs from a LSTM Seq-to-Seq model (S2S) (Rush et al., 2015), a Pointer-Generator Network (PGN) model (See et al., 2017), a Bottom-Up Summarization (BUS) model (Gehrmann et al., 2018), a Bert based Extractive-Abstractive model (BertSum) (Liu and Lapata, 2019) and a jointly pretrained transformer based encoder-decoder model BART (Lewis et al., 2020). For the XSum dataset, we collect model outputs from a Topic-Aware CNN Model (Narayan et al., 2018), a Pointer-Generator Network (PGN) model, a randomly initialized (TransS2S) (Vaswani et al., 2017) and one initialized with Bert-Base (BertS2S) (Devlin et al., 2019).[2] Details of the models used are provided in §A.1.

**Annotation Collection** Using the above model generated summaries, we collect human annotations from three independent annotators for 250 articles from each dataset (with a total of 1250 model outputs on CNN/DM and 1000 on XSum). We annotate each sentence of a summary to break the judgement of factuality into smaller units. We present sentences in the context of the entire summary to identify discourse errors spanning multiple sentences. Annotations are a two step process: for each sentence in the summary, the annotator first selects whether the sentence is factual, and if marked not factual, identifies the category of each error based on our typology. [3] A sentence can be annotated with more than one category of errors to account for multiple errors within a sentence. We conduct the annotation task on the Amazon Mechanical Turk (MTurk) platform. To achieve high quality crowd-sourced annotations, we build an intuitive interface[4] which combines:

1. **Clear Instructions:** We explain the annotation scheme without assuming linguistic knowledge and give several examples for each category.
2. **Training and Evaluation:** We setup training tutorials for first time users to train and provide feedback on the task. We also setup a qualification test which tests their understanding of our annotation scheme and require annotators to obtain >85% score to qualify. Further, we continuously evaluate annotators during the task against artificially generated factual errors to ensure continued high quality.
3. **Fair Pay and Bonus:** All workers are paid 50% more than the average American minimum wage. We offer bonuses for scores of 60% or above on the continuous evaluation, and for completing sets of 10 annotations.

Further details on our interface are added in §A.6

**Inter-Annotator Agreement:** We report inter-annotator agreement in terms of Fleiss Kappa $\kappa$ (Fleiss, 1971). Following Durmus et al. (2020), we report the percentage $p$ of annotators that agree with the majority class. Each datapoint in our dataset corresponds to a sentence in a summary. We compute agreement on all 4942 annotated sentences. On the annotation of whether a sentence is factual or not we obtain $\kappa = 0.58$, with $p = 91\%$ of annotators agreeing with the majority class. As a comparison, Durmus et al. (2020) reports $p = 76.7\%$ average agreement. When all three annotators agree that a sentence is not factual, we obtain $\kappa = 0.39$ with $p = 73.9\%$ of annotators agreeing with the majority class on the eight category annotation (seven categories of errors and "other") which indicate a moderate agreement.

**Agreement with Domain Expert:** We measure agreement between the majority class of the three

[2] As we use publicly available model outputs, the summaries across different datasets are from different models owing to their availability.

[3] We experimented with Likert scale evaluation of full summaries in a pilot study. Such an annotation would not provide precise information about where in the summary an error appears and also resulted in lower agreement. Hence, we opted for sentence level judgements.

[4] We make the interface available for future human annotations that follow our typology

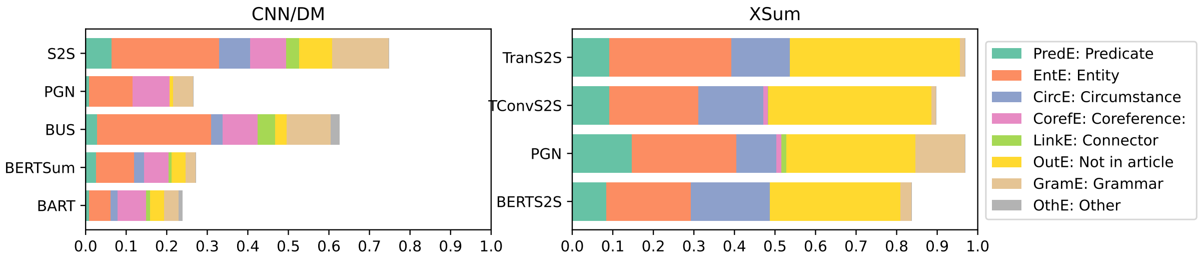


Figure 2: Proportion of summaries with factual errors based on collected annotations, with breakdown of the categories of errors within. Full specification of categories of errors in Table 1.

annotators and one expert annotator on 201 datapoints (10 summaries from CNN/DM and 10 summaries from XSum). We find a Kohen Kappa of $\kappa = 0.86$ indicating nearly perfect agreement. Previous work found agreement of $\kappa = 0.65$ between three crowd annotators and expert annotations of factuality (Falke et al., 2019). Even with more than nine workers, they report agreement with expert annotations of at most $\kappa = 0.74$. This improvement validates the robustness of our annotation interface and protocol which achieves higher agreement with fewer workers.

## 4 Summarization Model Analysis

We evaluate the performance of different summarization models in terms of factuality. Figure 2 visualizes the percentage of summaries with factual errors for each category model and dataset, with a breakdown of proportion of different error types within each. A summary is considered incorrect if it contains at least one sentence with a factual error. A sentence contains a factual error if the majority of annotators indicate the presence of an error (here we do not consider annotations where all three annotators disagree on the category).

**How factual are generated summaries across different datasets?** From our annotations, we observe that 60% of the summaries that were annotated contain at least one factual error. From Figure 2, we see that the XSum dataset has more factually incorrect model summaries (92%) than CNN/DM (43%). It poses more significant challenges in terms of factuality as all models produce $>$ 80% summaries with factual errors, with the best model (BertS2S) producing 83% wrong summaries. On the CNN/DM dataset, while state-of-the-art pretrained models like BERTSum and BART have better factuality numbers, the percentage of factually incorrect summaries is still high (23% for BERTSum and 27% for BART). The proportion of errors across different categories vary widely between the two datasets. For the CNN/DM dataset, the most frequent classes of errors are Entity Error (EntE) and Coreference Error (CorefE). For the XSum dataset they are Out of Article Error (OutE) and Entity Error (EntE). Note that there are no discourse errors (CorefE, LinkE) in the XSum dataset because the data only contains single sentence summaries. Additionally, we observe that OthE makes up a very small percentage ($\sim$ 1%) of errors overall showing that our typology is *complete* with most errors being mapped to one of our existing categories.

**How factual are generated summaries across different models?** From Figure 2, we observe that LSTM based models like S2S and BUS generate many incorrect summaries. Interestingly, PGN on CNN/DM has fewer summaries with factual errors (26%) compared to S2S (74%) and BUS (62%) potentially due to the extractive nature of CNN/DM and the copy based objective in PGN. PGN has been previously shown to produce highly extractive summaries on CNN/DM copying large portions of text (often entire sentences) (Gehrmann et al., 2018; Balachandran et al., 2021). On the more abstractive dataset XSum, PGN produces $>$ 96% factually incorrect summaries. We also observe that large-scale pretrained models improve factuality on both datasets, as also noted by Durmus et al. (2020), with more significant gains on CNN/DM. On CNN/DM, BERTSum and BART display half the error rate of BUS. In contrast, on XSum, BertS2S improves over non-pretrained models by $\sim$ 10% only, showing that XSum poses a significant challenge for factuality even in pretrained models.

Different models also exhibit different distributions in the error categories. LSTM based models have higher proportion of Grammatical Errors (GramE) while transformer and CNN based models have a lower proportion. For pretrained transformer models, we observe that the improved error-rate on the CNN/DM dataset can be attributed to improvements at the frame level (PredE, EntE, CircE) while the discourse level errors still remain a challenge. Errors CorefE, LinkE account for a higher proportion of errors in BERTSum and BART compared to the other models.

## 5 Factuality Metric Evaluation

We propose the FRANK dataset resulting from the human annotation study as a common benchmark to assess different factuality metrics. We provide an evaluation protocol of factuality metrics, which controls for dataset biases, and a fine grained analysis of the strengths of each metric.

### 5.1 Benchmark

The FRANK benchmark provides a diverse dataset for evaluating various metrics on their ability to capture factual errors. Notably, our benchmark has *factual error diversity*, as it covers all types of errors described in the typology in §2, and *data diversity* as it combines 2250 summaries from different systems and datasets. Our annotations go beyond binary labels of factuality on a summary by providing fine-grained category annotations for every sentence. This allows us to determine how well each metric can capture each type of error. Furthermore, through averaging of sentence level judgements, we can also obtain a factuality scores (0 to 1 range) for a summary. To measure the degree that automated metrics capture a certain characteristic, we compute their correlation with human judgements and report Pearson correlation and Spearman rank correlation along with their p-values.

We evaluate different classes of metrics against the FRANK benchmark. We select four general summarization metrics. ROUGE, BLEU, and Meteor are n-gram based metrics and computed with respect to the reference summary. BERTScore (Zhang et al., 2020) computes BERT (Devlin et al., 2019) contextual embeddings on summary and source article and measures distances between matched embeddings. We select five metrics focused on factuality. As Goodrich et al. (2019), we use a simple OpenIE (Banko et al., 2007) baseline. This involves extracting OpenIE triples and matching them through sentence embeddings (Reimers and Gurevych, 2019). FactCC (Kryscinski et al., 2020) and DAE (Goyal and Durrett, 2020) are entailment based metrics. FactCC operates with sentences as claims, while DAE uses dependency level entailment. FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020) are two question answering and generation metrics (QGA). More details on the differences between these metrics is in §A.2.

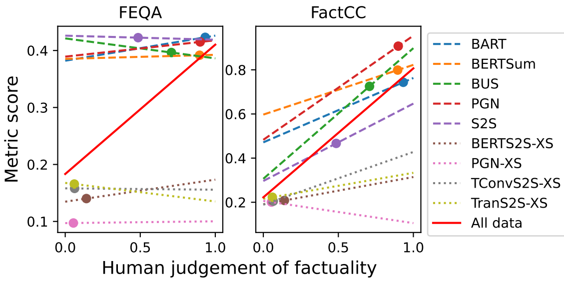


Figure 3: Correlation between metrics and human judgement on subsets of data. The $x$ and $y$ axis represent the human judgement the metric scores respectively. The red line is a linear regression fitted on full data. Each dotted line is a linear regression fitted on a model-dataset subset. Each colored point has coordinates equal to average factuality judgement, and metric score for its corresponding partition.

### 5.2 Controlling for Dataset Biases

Since our benchmark contains diverse summaries from different datasets and models, dataset biases can hamper accurate reporting. In Figure 3, we visually show correlations between two factuality metrics (FEQA and FactCC) and human judgement on the entire data and on partitions of the data. For both metrics, we notice that the slope (an unscaled measure of correlation) of the line fitted through the entire data (red line) is significantly larger. In FEQA, the dotted lines (fitted on subsets of the data of each model and dataset) are almost horizontal. This likely indicates the presence of a confounding variable associated with the properties of each system and dataset. This can lead to false measures of high correlation if not accounted for. To address this, we suggest to control for confounding variables using partial correlations. We include details on partial correlations in the Appendix. In this case, both the system and the dataset are taken to be confounding variables.

### 5.3 Results

In Table 2, we report the partial Pearson correlation and Spearman rank correlation coefficients with human judgements for each metric, along with their

| Metrics | All data | | | | CNN/DM | | | | XSum | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pearson | | Spearman | | Pearson | | Spearman | | Pearson | | Spearman | |
| | $\rho$ | p-val | $r$ | p-val | $\rho$ | p-val | $r$ | p-val | $\rho$ | p-val | $r$ | p-val |
| BLEU | 0.10 | 0.00 | 0.07 | 0.00 | 0.08 | 0.01 | 0.08 | 0.01 | 0.14 | 0.00 | 0.20 | 0.00 |
| METEOR | 0.14 | 0.00 | 0.11 | 0.00 | 0.12 | 0.00 | 0.10 | 0.00 | 0.15 | 0.00 | 0.10 | 0.00 |
| Rouge-1 | 0.14 | 0.00 | 0.10 | 0.00 | 0.12 | 0.00 | 0.10 | 0.00 | 0.15 | 0.00 | 0.09 | 0.01 |
| Rouge-2 | 0.12 | 0.00 | 0.08 | 0.00 | 0.08 | 0.00 | 0.07 | 0.01 | 0.17 | 0.00 | 0.14 | 0.00 |
| Rouge-L | 0.13 | 0.00 | 0.09 | 0.00 | 0.11 | 0.00 | 0.09 | 0.00 | **0.16** | 0.00 | 0.10 | 0.00 |
| OpenIE | 0.11 | 0.00 | 0.02 | 0.36 | 0.16 | 0.00 | 0.15 | 0.00 | 0.00 | 0.93 | -0.45 | 0.00 |
| BERTS P | -0.02 | 0.35 | -0.01 | 0.69 | 0.00 | 0.95 | -0.01 | 0.65 | -0.04 | 0.25 | 0.02 | 0.57 |
| BERTS R | -0.03 | 0.14 | -0.05 | 0.03 | -0.04 | 0.18 | -0.06 | 0.04 | -0.03 | 0.34 | 0.02 | 0.58 |
| BERTS F1 | -0.03 | 0.16 | -0.03 | 0.13 | -0.02 | 0.43 | -0.05 | 0.10 | -0.04 | 0.26 | 0.02 | 0.53 |
| FEQA | 0.00 | 0.83 | 0.01 | 0.60 | -0.01 | 0.76 | -0.01 | 0.72 | 0.02 | 0.45 | 0.07 | 0.04 |
| QAGS | 0.06 | 0.00 | 0.08 | 0.00 | 0.13 | 0.00 | 0.09 | 0.00 | -0.02 | 0.48 | 0.01 | 0.65 |
| DAE | 0.16 | 0.00 | 0.14 | 0.00 | 0.25 | 0.00 | 0.24 | 0.00 | 0.04 | 0.16 | 0.28 | 0.00 |
| FactCC | **0.20** | 0.00 | **0.30** | 0.00 | **0.36** | 0.00 | **0.33** | 0.00 | 0.07 | 0.02 | **0.25** | 0.00 |

Table 2: Partial Pearson correlation and Spearman rank correlation coefficients and p-values between human judgements and metrics scores. Comparisons should be made along with the pairwise Williams test found in Table 4.

$p$-values indicating statistical significance.

**How do different metrics correlate with human judgements?** From Table 2 we observe that all metrics exhibit low correlations with human judgements of factuality. The best metric overall is FactCC with 0.20 Pearson and 0.30 Spearman correlation. Interestingly, we observe that general summarization metrics BLEU, Rouge, and METEOR, and the OpenIE baseline have statistically significant correlations with factuality, close to FactCC ($\rho = 0.14$ for Rouge-1 and METEOR versus $\rho = 0.20$ for FactCC). The entailment metrics (FactCC and DAE) have the two highest correlations and are statistically significant. The two QGA metrics have lower overall correlation. FEQA's correlation is not statistically significant. QAGS has low, but significant correlation of $\rho = 0.06$.

**How well do different metrics capture errors in different datasets?** In Figure 4, we observe that entailment metrics have significantly higher partial Pearson correlation on the CNN/DM dataset than XSum where their correlation is reduced by a factor of four. QAGS and the OpenIE baseline have similar behavior. This suggests that these metrics capture the error types from CNN/DM better that those from XSum. Specifically, XSum has uniquely high Out of Article (OutE) errors which they might not capture well. This also highlights the importance of data diversity in building and benchmarking factuality metrics to avoid overfitting to certain types of errors.

**How well do different metrics capture errors from pretrained and non-pretrained models?** On the CNN/DM dataset we observe that entailment metrics and QAGS perform significantly better on non-pretrained models. This indicates that the artificial factual errors on which entailment metrics are trained on are closest to the mistakes that non-pretrained models make. This also suggests that the errors made by pretrained models might be more difficult to capture by these metrics. These trends are less clear on the XSum dataset which we again attribute to high Out of Article (OutE) errors in the pretrained and non-pretrained models (ref Figure 2)

### 5.4 Error Analysis

Figure 4 shows partial Pearson correlation on six subsets of the data. To understand capabilities of metrics across the broad categories of errors (semantic frame errors, discourse errors, and content verifiability errors) we perform an ablation study. For each category, we compute the variation in partial correlation with errors from that category omitted. In Figure 5, we visualize the influence of a given type of error using the variation for each metric and category. A higher positive bar indicates that the error type was a significant contributer to the overall correlation (or metric highly correlates with error) causing the correlation without it to

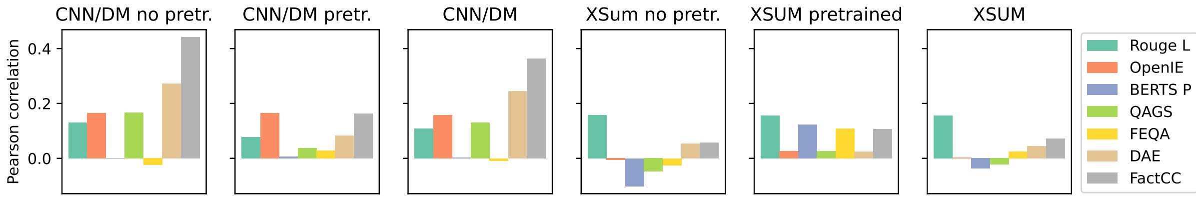


Figure 4: Partial Pearson correlation on different partitions of the data. Entailment metrics have highest correlation on pretrained models in the CNN/DM dataset. Their performance degrades significantly on XSum.

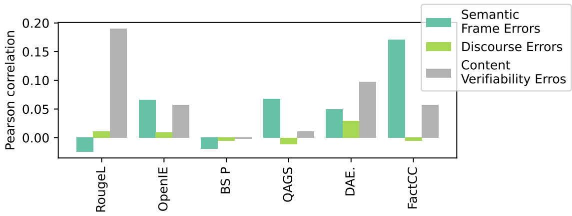


Figure 5: Variation in partial Pearson correlation when omitting error types. Higher variation indicates greater influence of an error type in the overall correlation.

drop.

**General Summarization metrics** Unsurprisingly, we observe that Rouge L is best correlated with content verifiability errors (which contains Out of Article Errors) as n-gram matches detect them. Rouge L has negative correlation with semantic frame errors and low correlation with discourse level errors indicating that n-gram matching fails to capture them. We observe that OpenIE is more correlated with semantic frame errors. The metric matches entities and verifies the predicate that relates them and hence is able to capture semantic frame errors. BertScore has low correlation overall, being more correlated with content verifiability errors and negatively correlated with discourse errors.

**QGA metrics** Both QGA metrics have negative correlation with discourse errors suggesting that QGA metrics are not able to capture coreference errors or discourse link errors potentially due to the entity oriented questions in their training data. FEQA additionally is also negatively correlated with semantic frame errors and has low positive correlation with content verifiability errors. In contrast QAGS is best correlated with semantic frame errors.

**Entailment metrics** Both entailment metrics correlate well with semantic frame and content verifiability errors. DAE has the highest correlation of all metrics with discourse errors suggesting that entailment at the dependency level can help model discourse errors (CorefE and LinkE). FactCC is nearly uncorrelated in this category, indicating that artificially generated factual errors need to go beyond simple pronoun swaps to train models to capture discourse errors. FactCC had best overall partial correlation which can be attributed to FactCC being able to capture semantic frame and content verifiability errors well.

## 6 Related Work

Kryscinski et al. (2019) and Fabbri et al. (2020) find that standard n-gram based metrics have low correlation with human judgements of factuality. Motivated by this, several automated metrics falling in two paradigms were proposed to improve the evaluation of factuality.

**Entailment Classification** Goodrich et al. (2019); Kryscinski et al. (2020); Maynez et al. (2020); Goyal and Durrett (2020) model factuality as entailment classification breaking down the summary into smaller units, such as sentences, which are verified against the original article. However, modeling factuality as a classification task requires supervision on factual and hallucinated data. FactCC (Kryscinski et al., 2020) is trained on the CNN/DM dataset augmented with four types of artificial mistakes as supervision.

**Question Generation and Answering (QGA)** FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020) are two metrics which reduce factuality evaluation to question generation and answering. These methods use a question generation model to obtain questions from the output summary and a question answering model to answer them, separately using the article and the output summary.

**Prior Efforts on Factuality Annotations of Summaries** Fabbri et al. (2020) and Maynez et al.

(2020) have collected annotations on the CNN/DM and XSum dataset respectively. In this work we cover both datasets to ensure greater data diversity. Other efforts (Kryscinski et al., 2020; Wang et al., 2020; Durmus et al., 2020) were smaller in scale Durmus et al. (2020) and Kryscinski et al. (2020) annotated 200 and 503 sentences while Wang et al. (2020) annotated 470 summaries (we collect judgements on 2250 summaries). Crucially, all previous efforts portray factuality as a binary label without variations in degree or type of factual errors.

## 7 Conclusion

In this work we provide a linguistically grounded typology of factual errors which we use to collect FRANK, a dataset of human annotations of 2250 summaries covering both CNN/DM and XSum datasets. We use FRANK to assess the factuality of summarization systems and benchmark recently proposed factuality metrics highlighting the types of errors they can capture. With the FRANK benchmark we have started moving away from a summary-level binary understanding of factuality.

## 8 Ethical Considerations

We have collected crowd annotations using the Amazon Mechanical Turk platform. Workers were paid 50% more than the average American minimum wage and offered additional bonuses as an incentive to maintain high quality work. No information about the workers will be released and worker IDs will be anonymized.


## Acknowledgements

The authors are grateful to the anonymous reviewers for their feedback, and to Anjalie Field, Rishabh Joshi, Alissa Ostapenko, Dheeraj Rajagopal, Evangelia Spiliopoulou, Shuly Wintner, and the members of the Tsvetshop group for their invaluable feedback and support in various stages of the project. This material is based upon work supported by the DARPA CMO under Contract No. HR001120C0124, and in part by the National Science Foundation under Grants No. IIS2040926 and No. IIS2007960. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily state or reflect those of the United States Government or any agency thereof.


## References


Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, and Yulia Tsvetkov. 2021. StructSum: Summarization via structured representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2575–2585, Online. Association for Computational Linguistics.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gillian Brown and G. Yule. 1983. *Discourse Analysis*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

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

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

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.

Charles J Fillmore et al. 1976. Frame semantics and the nature of language. In *Proceedings of the Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32. New York.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 166–175. ACM.

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.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

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

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

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

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.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

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.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

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

Roser Sauri. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University, USA.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

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.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. 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 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

# A Appendices

## A.1 Model details

We provide details of the models used in the human evaluation task to construct FRANK.

### A.1.1 CNN/DM datset

On the CNN/DM (Hermann et al., 2015) dataset we use five different models. We use the preprocessed model outputs provided by Fabbri et al. (2020).
**S2S** an LSTM based Sequence-to-Sequence with attention model (Rush et al., 2015)
**PGN** an LSTM based Pointer-Generator Network with Copy Mechanism (See et al., 2017)
**BUS** Bottom-Up Summarization (Gehrmann et al., 2018) - a Pointer-Generator model with a data-efficient content selector to over-determine phrases in a source document that should be part of the summary.
**BERTSum** summarization with pretrained encoders (Liu and Lapata, 2019)
**BART** (Lewis et al., 2020)

### A.1.2 XSum dataset

On the XSum dataset (Narayan et al., 2018) we use four different models. All model outputs for this dataset are taken from (Maynez et al., 2020)
**PGN** pointer-generator network from above (See et al., 2017)
**TConvS2S** Topic-Aware Convolution Sequence-to-Sequence (Narayan et al., 2018)
**TranS2S** A randomly initialized Transformer (Vaswani et al., 2017) encoder-decoder model fine-tuned on the XSum dataset
**BERTS2S** Transformer encoder-decoder model with parameter sharing (Rothe et al., 2020) where both encoder and decoder are initialized with the BERT-Base checkpoints (Devlin et al., 2019) and fine-tuned on XSum

## A.2 Metrics

In this work we compare the following five metrics.

**BERTScore (Zhang et al., 2020):** We report BERTScore Precision, Recall, and F1 between the model output and the reference summary. Our experiments show that recall and F1 do not correlate as well with the human judgement of factuality for BERTScore.

**OpenIE** : We use a simple baseline based on OpenIE (Banko et al., 2007) and Sentence-BERT (Reimers and Gurevych, 2019). We use OpenIE (Banko et al., 2007) to extract subject-relation-object triplets from the article, reference summary, and model generated summary. We consider binary relations only and thus use the first two arguments of the relation.[5] After replacing corefering entity mentions with the main mention of the cluster[6], we use BERT base Sentence-BERT (Reimers and Gurevych, 2019) to obtain embeddings of each element of the subject-relation-object triplets extracted by OpenIE. Two relation triplets are considered to be equivalent if their embeddings have cosine similarity higher than a threshold for all three elements of the triplet (we use 0.6 as threshold after a grid search between 0.5 and 0.9 on data from our pilot study).

**FEQA (Durmus et al., 2020):** FEQA is a question generation and answering (QGA) factuality metric. We relied on the original implementation of the authors for this metric as well as their pretrained model weights. We used the full summary to generate questions and we answer them both using the summary and article text.

**QAGS (Wang et al., 2020):** QAGS is another QGA metric. The authors kindly provided outputs on the FRANK benchmark generating 10 questions for each summary.

**DAE (Goyal and Durrett, 2020):** DAE is an entailment classification metric that operates on dependencies. The authors kindly provided outputs on the FRANK benchmark. We note that the model was trained with a max length of 128 after concatenating both article and summary. The CNN/DM articles can be significantly longer, thus the results reported for this metric involve truncating parts of the article.

**FactCC (Kryscinski et al., 2020):** FactCC is an entailment classification metric. We use the sentences of the model generated summary as input claims to the entailment classifier FactCC. For each sentence we obatain a binary factuality label. We take the average of these labels as the factuality score for the summary.

[5] We use the model and implementation from (Stanovsky et al., 2018) for OpenIE extraction.
[6] https://github.com/huggingface/neuralcoref

### A.3 Summarization System Analysis Details

See Table 1 for more details.

### A.4 Hotelling Williams Test

The correlation numbers in Table 2 should be read in combination with the pairwise Hotelling-Williams test Graham (2015) results in Table 4. The highlighted numbers indicate pairs of models for which the difference in correlation is statistically significant. We use partial correlations to run the test and compute metric-metric correlations.

### A.5 Mutual Exclusiveness of typology:

To understand if our annotations are mutually exclusive, we study cases where two annotators agree on the error category (majority class) and one disagrees (minority class). In Figure 6, we report the confusion between majority and minority classes. For each category as majority, we report the distribution of other categories as minority.

We observe that all categories with the exception of **OutE** are frequently confused with **NE** which stands for no factual error. This primarily due to the noise in the annotations collected by crowd workers. However, for category **CorefE** (coreference errors) the confusion is significantly higher with 69.7%. We have noticed the same trend in practice tutorials: crowd annotators easily overlook situations where the correct pronoun is used (in terms of number and gender) but no antecedent appears in the summary. Intuitively after reading the article, unless paying particular attention, it is easy to subconsciously associate referring expressions with entities in the article without noticing their absence in the summary. The error persists despite stating the scenario explicitly in the instructions. This indicates an issue with annotators rather than annotation scheme.

The other trend that we observe is that categories **PredE** (wrong relation) and **CircE** (wrong modifier) are often confused with **OutE** (outside information). In our definition of **OutE**, outside information corresponds to the presence of entities not mentioned in the article or relations that cannot be verified based on the article. The confusion with **PredE** indicates that annotators can have different judgements on whether a relation is verifiable based on the article. Similarly, but to a lesser degree, wrong circumstantial information might be considered unverifiable given the article.

Finally, there were relatively few discourse context errors **LinkE**, so the analysis is less statistically significant. Discourse context errors correspond to using a wrong connectors between different facts, for example different logical links. These were confused with **PredE** and **EntE** (wrong relation). The distinction between the two errors lies in the confusion between what an entity and a fact are, since **PredE** occurs at the frame level while **LinkE** at the discourse level. Note, that there was no confusion in the other direction (**PredE** being confused with **LinkE**).

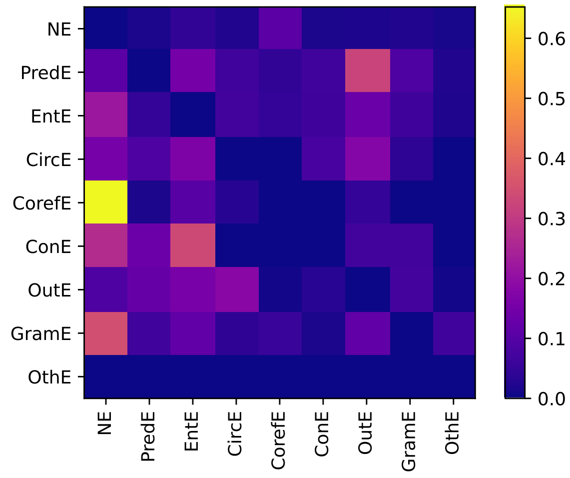


Figure 6: Confusion matrix of different types of errors. Entry at row $i$, column $j$ corresponds to the frequency of annotations that have F$i$ as the majority class and for which disagreeing annotator selected F$j$.

### A.6 Annotation Setup Details

Below are more details on the annotation set up.

**Clear Instructions** We explain the annotation scheme without assuming linguistic knowledge and give several examples for each category. We also provide a practival ste-by-step to determine the category of the errors.

**Training** Every first-time user has to go through a tutorial which exercises the comprehension of the annotation scheme. The tutorial presents an article and several hand-crafted summaries of the article that need to be annotated. It is designed to be very similar to the actual annotation task and to contain at least one occurrence of each category of error. Feedback is provided when a user selects the wrong category of error. This tutorial is not used to evaluate users, only to help them understand the different categories in a practical setting.

| | Incorrect | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 |
|---|---|---|---|---|---|---|---|---|---|
| Seq2Seq | 74.8% | 11% | 46% | 13% | 15% | 5% | 14% | 24% | 0% |
| PGN | 26.5% | 4% | 46% | 0% | 39% | 0% | 4% | 21% | 0% |
| Bottom Up | 62.6% | 6% | 56% | 6% | 17% | 9% | 6% | 21% | 4% |
| BERTSum | 27.2% | 10% | 37% | 10% | 23% | 3% | 13% | 10% | 0% |
| BART | 23.8% | 4% | 25% | 8% | 33% | 4% | 17% | 17% | 4% |
| PGN | 96.9% | 16% | 28% | 11% | 1% | 1% | 34% | 13% | 0% |
| TConvS2s | 89.8% | 10% | 24% | 18% | 1% | 0% | 45% | 1% | 0% |
| TranS2S | 96.9% | 10% | 32% | 15% | 0% | 0% | 44% | 1% | 0% |
| BERTS2S | 83.7% | 10% | 25% | 23% | 0% | 0% | 38% | 3% | 0% |
| All models | 60.0% | 10% | 36% | 13% | 10% | 3% | 27% | 12% | 1% |

Table 3: Proportion of summaries that include factual errors, with breakdown of the categories of errors according to our human study. F8 corresponds to errors that are not captured by our typology. Full specification of categories of errors in Table 1.

| | B | MET | R-1 | R-L | BS-P | OpIE | FEQA | QAGS | DAE | FCC |
|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | - | 0.83 | 0.77 | 0.85 | 0.04 | 0.26 | 0.03 | -0.01 | 0.05 | 0.06 |
| METEOR | 0.83 | - | 0.87 | 0.85 | 0.04 | 0.28 | 0.02 | -0.02 | 0.08 | 0.07 |
| Rouge-1 | 0.77 | 0.87 | - | 0.89 | 0.04 | 0.21 | 0.01 | -0.03 | 0.09 | 0.07 |
| Rouge-L | 0.85 | 0.85 | 0.89 | - | 0.03 | 0.21 | 0.01 | -0.04 | 0.08 | 0.07 |
| BERTS P | 0.04 | 0.04 | 0.04 | 0.03 | - | 0.00 | 0.00 | 0.02 | -0.02 | -0.04 |
| OpenIE | 0.26 | 0.28 | 0.21 | 0.21 | 0.00 | - | -0.01 | 0.09 | 0.10 | 0.15 |
| FEQA | 0.03 | 0.02 | 0.01 | 0.01 | 0.00 | -0.01 | - | -0.01 | 0.03 | 0.04 |
| QAGS | -0.01 | -0.02 | -0.03 | -0.04 | 0.02 | 0.09 | -0.01 | - | 0.08 | 0.09 |
| DAE | 0.05 | 0.08 | 0.09 | 0.08 | -0.02 | 0.10 | 0.03 | 0.08 | - | 0.10 |
| FactCC | 0.06 | 0.07 | 0.07 | 0.07 | -0.04 | 0.15 | 0.04 | 0.09 | 0.10 | - |

Table 4: Pearson correlation between metrics. If value is in green, the metrics are not the same significant to the 0.05 threshold with the Hotelling Williams test.

**Qualification test** To participate in the annotation, users have to obtain a minimum score of 85% on a qualification test. The test comprehends an article and several summaries to be annotated. It contains at least one instance of each category of error. We use this test to verify that users can effectively recognize error categories. This ensures that users are able to perform the task correctly, but does not enforce that high standards of work quality are maintained throughout the annotation task.

**Continuous evaluation** We continuously evaluate a user by verifying that they read the text. For every article that is annotated, we ask to identify one of three entities that was not present in the article. We also monitor the annotations on artificially altered sentences that are randomly inserted at the end of summaries. Wrong sentences contain one of the following errors: negation of declarative sentences (PredE), pronoun swap (CorefE), sample sentence from another article (OutE), word scrambling (GramE). We immediately block users that fail the entity test or perform poorly on these sentences (less than 50% of correct answers on altered sentences) to ensure high quality annotations.

**Bonuses** All workers are paid 50% more than the average American minimum wage but we offer bonuses for scores of 60% or above on the continuous evaluation, and for completion a sequences of 10 annotations. We observe that bonuses increase the percentage of users with high continuous evaluation scores (<10% blocked users with bonuses versus 30% without bonuses).

### A.7 Correlation with Confounding Variables

Partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed. Although we are unaware of the exact confounding variable, we use the categorical variable $C$ of which system and dataset the summary was generated from.

Let $M_k$ represent the output of metric $k$ on the summaries. To compute partial correlation between $M_k$ and human judgements $H$ which we treat as random variables, we solve the two regression problems $M_k|C = c \sim w_{M_k}c$ and $H|C = c \sim w_H c$ and get the residuals:

$$\Delta M_k = M_k - \hat{w}_{M_k} C$$

$$\Delta H = M_k - \hat{w}_H C$$

And then calculate the correlation between these residuals $\rho(\Delta M_k, \Delta H)$ instead of the original random variables. Since partial correlations are proper correlations between random variables, we can apply statistical significance tests without any modification.

### A.8 Annotation Interface

We include screenshots of the annotation interface which we will make available.

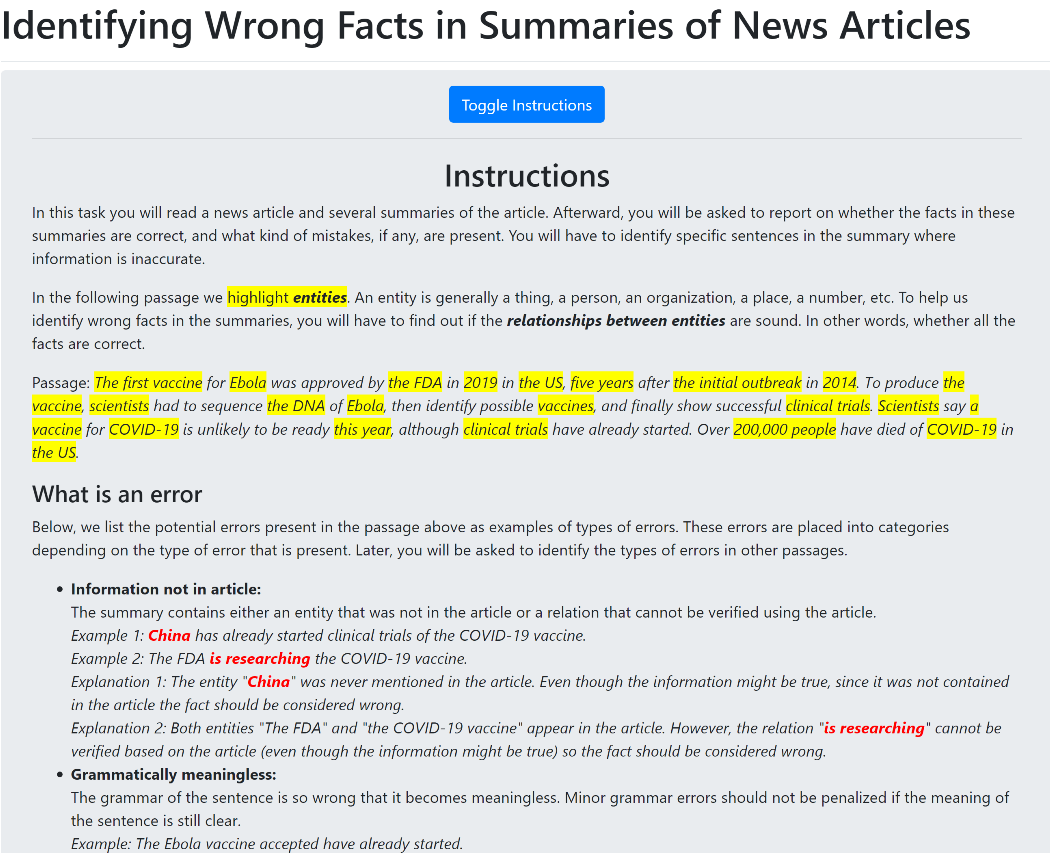


Figure 7: Instructions can be toggled.

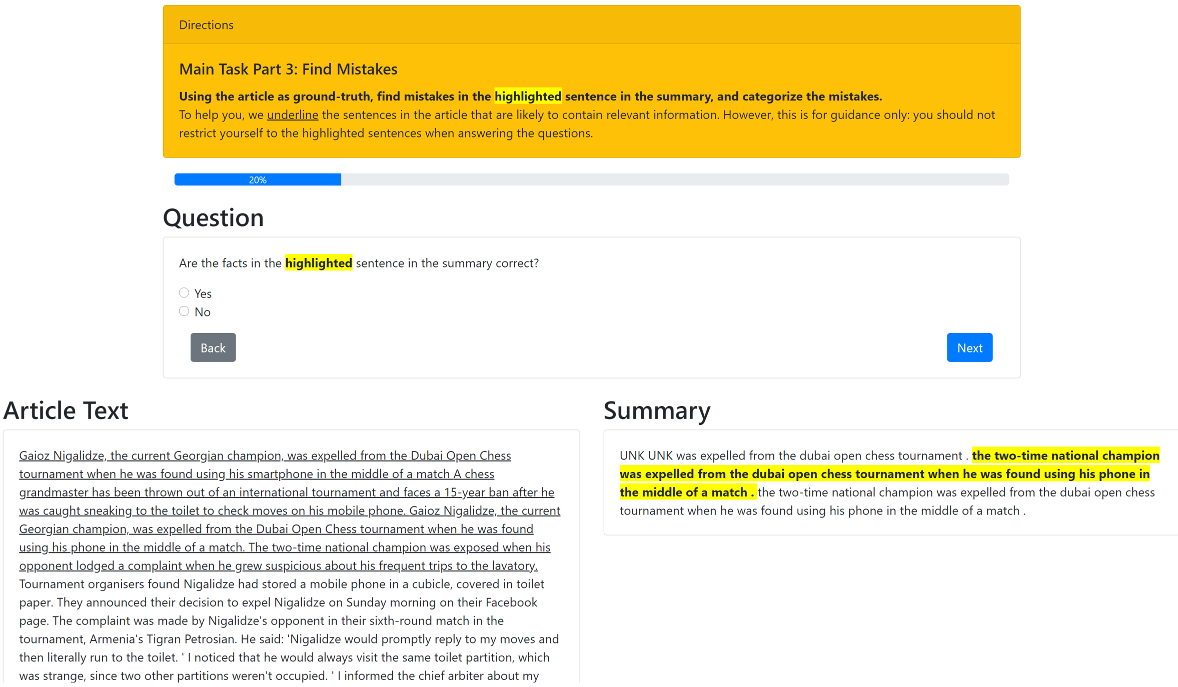


Figure 8: The sentences being annotated is highlighted in yellow. Relevant text is underlined in the article plain text.

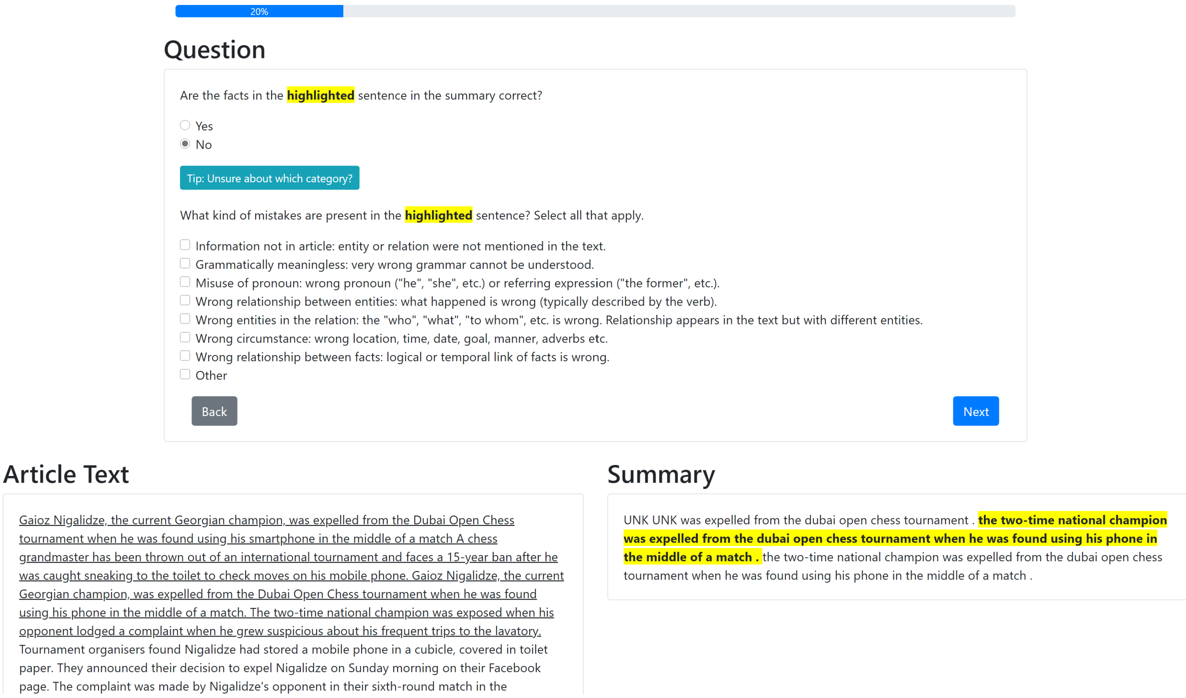


Figure 9: After selecting that the sentence is not factual annotators choose the category of error.


Figure 10: Articles web pages are provided.

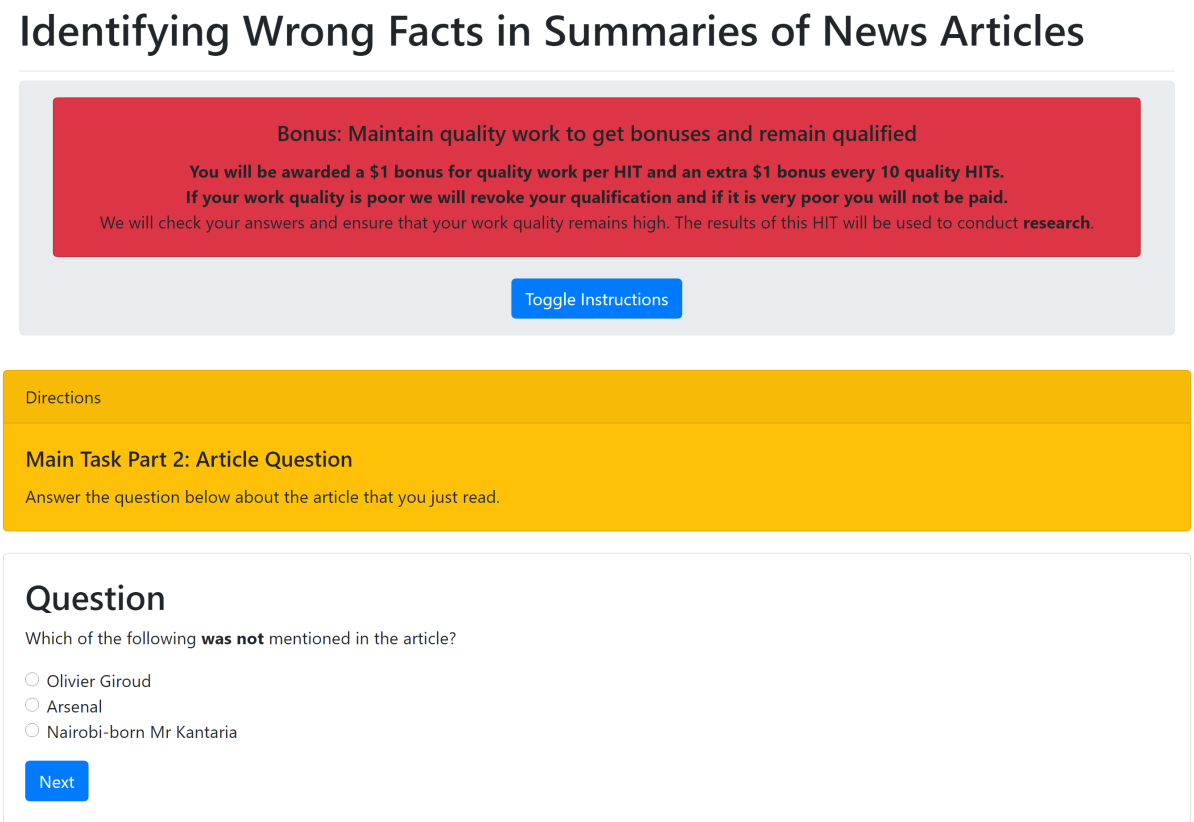


Figure 11: Entity question to ensure annotators read the text.