Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

Measuring Faithfulness of Abstractive Summaries

Tim Fischer **Steffen Remus** **Chris Biemann**

Language Technology Group
Department of Informatics
Universität Hamburg, Germany
{firstname.lastname}@uni-hamburg.de

Abstract

Recent abstractive summarization systems fail to generate factually consistent – faithful – summaries, which heavily limits their practical application. Commonly, these models tend to mix concepts from the source or hallucinate new content, completely ignoring the source. Addressing the faithfulness problem is perhaps the most critical challenge for current abstractive summarization systems. First automatic faithfulness metrics were proposed, but we argue that existing methods do not yet utilize the full potential that this field has to offer and introduce new approaches to assess factual correctness. We evaluate existing and our proposed methods by correlating them with human judgements and find that BERTScore works well. Finally, we conduct a qualitative and quantitative error analysis, which reveals common problems and indicates means to further improve the metrics.

1 Introduction

Abstractive summarization is the task of generating an informative and fluent summary that is faithful to the source document. Recent progress in neural text generation has led to significant improvements and well-performing state-of-the-art abstractive summarization systems (Zhang et al., 2019; Lewis et al., 2020). Despite these advances, recent models fail to meet one of the essential requirements of practical summarization systems: information of a generated summary must match the facts of the source document. We follow Cao et al. (2018) and refer to this aspect as faithfulness in this work. Recent studies have shown that around 30% of automatically generated summaries from neural summarization systems contain unfaithful information (Cao et al., 2018; Falke et al., 2019; Kryscinski et al., 2019), especially when a sentence combines content from multiple source sentences (Lebanoff et al., 2019). Table 1 shows a misleading and unfaithful summary demonstrating this issue.

Source	The restaurant began serving puppy platters after a new law was introduced allowing dogs to eat at restaurants – as long as they were outdoors!
Summary	New rules have come into place that you can eat your dog.

Table 1: A generated, unfaithful summary found in the XSUM hallucination dataset by Maynez et al. (2020).

Researchers identified multiple challenges for developing faithful systems. One challenge is evaluation, as current automatic metrics are inadequate. Typical metrics like ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005) are insensitive to semantic errors. These n-gram-based approaches weight all portions of the text equally, even when only a small fraction of the n-grams carry most of the semantic content. Consequently, factual inconsistencies caused by small changes are overshadowed by high n-gram overlaps. Another challenge is the optimization of abstractive models. Generating summaries that highly overlap with human references does not guarantee faithful summaries (Zhang et al., 2020b).

Initial work on metrics to automatically assess faithfulness will be discussed in Section 2 and 3, however, no consensus has been reached to date. We argue that the currently available means to automatically evaluate faithfulness do not use the full potential that current NLP methods offer. In this work, we explore new methods to assess the faithfulness of generated texts and compare them to existing approaches. Finally, we perform a qualitative and quantitative error analysis by investigating the outputs of all methods to analyze their problems and to reveal ways to improve them. We study the following research questions (RQs) in this work:

1. Which faithfulness metric correlates best with human judgements?
2. What are problems of faithfulness metrics and how can we address them?

Together with this work, we release an open-source Python library[1] that allows reproduction of our results and utilization of all discussed metrics by others to evaluate faithfulness.

2 Related Work

The lack of automatic evaluation metrics for faithfulness has motivated researches to develop new metrics that ideally mimic human judgements of factual consistency. Popular approaches are based on question answering (Wang et al., 2020; Durmus et al., 2020), textual entailment (Falke et al., 2019; Maynez et al., 2020) and contextual embeddings (Kryscinski et al., 2020).

Nan et al. (2021) focus on the problem of unfaithful entities where model-generated summaries contain named entities that do not appear in the source document. The authors perform named entity recognition and calculate the percentage of entities in the summary that can be found in the source. A low percentage means entity hallucination is severe. In addition, they propose precision-target and recall-target, which capture the entity-level accuracy of the generated summary with respect to the ground truth summary.

Goodrich et al. (2019) propose to measure the factual correctness with relation extraction methods. Facts are represented as subject-predicate-object triples and faithfulness is defined as the precision between the facts extracted from the generated summary and target summary.

3 Methods

We re-implement and modify popular faithfulness metrics as well as propose new methods (SentSim, NER, SRL) that extract and compare different information from text to assess factual consistency.

3.1 BERTScore

BERTScore (Zhang et al., 2020a) is an automatic evaluation metric for text generation. It utilizes contextual embeddings to compute a similarity score between every token in the candidate sentence and reference sentence. Computing the similarity with contextual embeddings is effective for matching paraphrases as well as capturing distant dependencies and ordering.

Let x be a reference sentence $x = x_1, ..., x_n$ and a y be candidate sentence $y = y_1, ..., y_m$ tokenized into tokens x_i and y_j, respectively. An embedding model maps theses sentences to two sequence of vectors $\mathbf{x_1}, ..., \mathbf{x_n}$ and $\mathbf{y_1}, ..., \mathbf{y_m}$. Every token in x is matched to a token in y to compute recall and each token in y is matched to a token in x to compute precision using maximum matching: each token is aligned to the most similar token in the other sentence. Three variants of BERTScore (precision, recall, F1) are shown below:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{y_j \in y} \mathbf{x_i}^T \mathbf{y_j}$$

$$P_{BERT} = \frac{1}{|y|} \sum_{y_j \in y} \max_{x_i \in x} \mathbf{x_i}^T \mathbf{y_j}$$

$$F1_{BERT} = 2 \frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}}$$

We optimize BERTScore by selecting layer 8 of RoBERTa-large (Liu et al., 2019) fine-tuned on Multi-NLI (Williams et al., 2018) (roberta-large-mnli on Hugging Face) to compute embeddings.

3.2 Textual Entailment (TE)

Textual Entailment (Dagan et al., 2005) is a popular approach to measure factual consistency employed e.g. by Falke et al. (2019), Maynez et al. (2020), Durmus et al. (2020). The basic intuition is that all information in a summary should ideally be entailed by the source document or perhaps be neutral to the source document, but the summary should never contradict it.

Let E be a TE model that predicts the probability $E(a, b)$ that text b is entailed by text a. The faithfulness score f of a summary S consisting of sentences $s_1, ..., s_n$ with respect to the original document D with sentences $d \in D$ can be computed in 3 different ways:

$$f_{s2s}(S) = \frac{1}{n} \sum_{i=1}^{n} max_{d \in D} E(d, s_i)$$

$$f_{d2s}(S) = \frac{1}{n} \sum_{i=1}^{n} E(D, s_i)$$

$$f_{top2s}(S) = \frac{1}{n} \sum_{i=1}^{n} E(P, s_i)$$

The sentence-to-sentence (s2s) scoring method checks if every summary sentence is entailed by any source sentence. The document-to-sentence (d2s) checks if every summary sentence is entailed by the source document. The top-to-sentence (t2s) checks if every summary sentence is entailed by

[1]https://github.com/bigabig/faithfulness

the k (=3) most similar source sentences (calculated by comparing cosine-similarities of sentence embeddings) forming paragraph P.

We use BART-large (Lewis et al., 2020) and RoBERTa-large fine-tuned on Multi-NLI in our experiments to compute entailment and sentence-transformers[2] to compute sentence embeddings (for t2s).

3.3 Question Generation & Question Answering (QGQA)

The QGQA framework was introduced by Durmus et al. (2020) and Wang et al. (2020) and has been used in follow-up work, e.g. Maynez et al. (2020); Dong et al. (2020). The basic intuition of this framework is: if we ask questions about a summary and its source, we expect to receive similar answers if the summary is faithful. Naturally, more matched answers imply a more faithful summary as the information addressed by these questions is consistent between summary and source.

QGQA framework performs the following steps to detect factual inconsistencies:

1. An answer candidate selection (AS) model selects important text spans from the summary.
2. A question generation (QG) model generates a set of question about the summary using the answer candidates.
3. A question answering (QA) model answers these questions using both the source document and the generated text.
4. The faithfulness score is computed based on the similarity of the corresponding answers.

A similarity metric is necessary to compare corresponding answers. We empirically find $F1$ surface (token-level) similarity performs best (Appendix A.1).

We use the transformers library (Wolf et al., 2020) to implement this framework. Named entities and noun phrases are extracted with spaCy[3] as answer candidates. We use T5-base[4] as QG model to generate 5 questions per candidate, but filter out duplicates, bad questions (questions that cannot be answered by QA model given the summary) and low probability questions to have at most 10 questions per summary. RoBERTa-large fine-tuned on SQUAD2 (Rajpurkar et al., 2018) is used as QA model (deepset/roberta-large-squad2 on Hugging Face).

[2] all-mpnet-base-v2 from https://www.sbert.net/index.html

[3] en_core_web_lg from https://spacy.io/

[4] https://github.com/fajri91/question_generation

3.4 Sentence Similarity (SentSim)

The intuition of SentSim to measure faithfulness is that the information expressed in the summary should be the same as in the source document but paraphrased. Therefore, a summary sentence should be very similar to one or multiple important source sentences.

Abstractive summaries are written using different wordings and formulations to express the same information. Consequently, SentSim has to successfully deal with highly paraphrased text detecting similar concepts expressed with different words on the one hand. On the other hand, it has to differentiate between similar and contrasting or contradicting information so that it can actually be used to score faithfulness.

We propose the following strategy to asses faithfulness with sentence similarity:

1. Apply sentence splitting to the source document and summary to obtain lists of sentences.
2. Match every summary sentence with the most similar source sentence to compute precision; vice-versa to compute recall.

The precision variant (recall is analog, F1 as usual) of SentSim is defined as follows: let $S = \{s_1, s_2, ..., s_N\}$ be the set of summary sentences and let $D = \{d_1, d_2, ..., d_M\}$ be the set of document sentences, then

$$P_{SentSim} = \frac{1}{|S|} \sum_{s_j \in S} \max_{d_i \in D} sim(d_i, s_j)$$

We utilize spaCy to apply sentence splitting and experiment with various implementations of $sim()$. We empirically find that F1 and BERTScore perform well to score and align sentences (Appendix A.1).

3.5 Named Entity Recognition (NER)

Factual inconsistencies can occur at different levels. The entity hallucination problem occurs when a summary contains named entities that do not appear in the source document. Intuitively, a summary containing many entities that do not appear in the source is less faithful than a summary that contains the same entities as the source.

We propose the following strategy to calculate faithfulness with NER:

1. Identify entities in summary and source.
2. Group entities by their label (e.g. PER).
3. For each summary entity, calculate the most similar entity of the same group in the source document and its similarity score.
4. The faithfulness score is the average over all similarity scores.

We use spaCy to extract named entities and empirically find that Exact Match and F1 perform well to compare them (Appendix A.1). Please note, this approach does not capture other aspects that influence faithfulness like relations between entities or context surrounding entities.

3.6 Open Information Extraction (Open IE)

At relation level, we compare the relations between entities appearing in the source document and the summary. The relation hallucination problem occurs when a summary contains the same entities as the source document but their relations do not appear in the source document.

Naturally, if a summary contains many relations not present in the source document it is less faithful than a summary that contains the same relations. More matched relations imply a more faithful summary since not only the entities but also their interaction is consistent. In contrast to NER, a perfect match of summary relations with source relations can guarantee a faithful summary.

We propose the following strategy to calculate faithfulness with Open IE:

1. Apply a co-reference resolution system to replace all pronouns in the texts with their respective entity.
2. Apply an Open IE system to extract summary triples ($R(s)$) and source triples ($R(d)$) of the form (subject, relation, object) representing any fact in the given text.
3. Compute a faithfulness score based on the comparison of the extracted relations.

We use the Stanford CoreNLP toolkit for Open IE (Angeli et al., 2015), which includes an option to apply co-reference resolution as pre-processing step. We experiment with different methods to compare triples. The Relation Matching Rate (Zhu et al., 2021) operates on fact triples and basically measures the ratio of correct hits. Additionally, we linearize fact triples by concatenating the subject, relation and object to measure similarity with typical metrics. We empirically find that F1 or BERTScore work best (Appendix A.1).

3.7 Semantic Role Labeling (SRL)

This approach is inspired by the YiSi metric (Lo, 2019). YiSi measures similarity between two sentences by aggregating the semantic similarities of semantic structures. We argue that comparing semantic frames in contrast to comparing tokens as e.g. in BERTScore brings more linguistic structure into the faithfulness assessment. This process can find crucial differences between the argument structure of summary and source, which is a desirable property considering faithfulness. It verifies whether summary phrases are used in a semantically similar way as in the source document and should help to identify cases where the summary differs from the originally intended meaning.

We propose the following strategy to calculate faithfulness with SRL:

1. Apply a SRL model to the summary and source document to obtain labeled phrases.
2. Optionally, filter and merge semantic role labels to increase robustness.
3. Group phrases by their label.
4. Align (a) source and summary phrases with same label using a similarity metric.
5. Aggregate the similarity scores of aligned phrases and average over all labels to compute faithfulness (f).

Formally, this calculation can be denoted as

$$\begin{aligned} a_{recall}(l) &= \frac{1}{|P_{S,l}|} \sum_{p_i \in P_{S,l}} \max_{p_j \in P_{D,l}} sim(p_i, p_j) \\ f_{metric} &= \frac{1}{|L|} \sum_{l \in L} a_{metric}(l) \end{aligned}$$

where $metric \in \{precision, recall, F1\}$. The precision variant of alignment (a) is analog to a_{recall}, F1 is calculated as usual. L is the set of all semantic labels, sim is a similarity metric comparing two texts, $P_{D,l}$ and $P_{S,l}$ are sets of phrases with label $l \in L$ for source D and summary S.

We use SRL BERT (Shi and Lin, 2019) of AllenNLP (Gardner et al., 2018) toolkit trained on the English OntoNotes 5 dataset (Hovy et al., 2006) for semantic role labeling. Following Lo (2019), we merge semantic role labels into more general role types (who, what, whom, when, where, why, how) for more robust performance. We empirically find computing similarity scores of phrases ($sim()$) works best with cosine-similarity (Appendix A.1).

4 RQ1: Best faithfulness metrics

We evaluate all faithfulness metrics described in Section 3 on the XSUM hallucination dataset (Maynez et al., 2020) as well as the SummEval dataset (Fabbri et al., 2021) and compute the correlation with human judgements. XSUM contains human faithfulness judgements (averaged to faithfulness scores) for 2000 document-summary pairs obtained by randomly sampling 500 articles from the XSUM (Narayan et al., 2018) test set and applying four different summarization models. Three annotators per document-summary pair were given the task to identify unfaithful text spans (hallucination spans) in the summary. The faithfulness score is roughly equivalent to the number of faithful words divided by number of total words of a summary. SummEval contains human faithfulness judgements for 1600 document-summary pairs obtained by randomly sampling 100 articles from the CNN/DailyMail (Hermann et al., 2015) test set and applying 16 different neural summarization models. Five crowd-sourced and 3 expert annotators were given the task to rate the factual consistency on a Likert scale from 1 to 5.

We apply a faithfulness metric on all document-summary pairs and calculate Spearman correlation (p) and Pearson correlation (r) coefficients between human judgements and predicted faithfulness scores. Results are reported in Table 2.

On the XSUM dataset, BERTScore achieves the highest correlation with human judgements. Entailment, SentSim and SRL perform similarly. On the SummEval dataset, SentSim and Entailment achieve the best correlation with human judgements. Open IE is last in both rankings.

Comparing XSUM and SummEval, there is a huge performance difference. This reason is twofold: First, we developed and optimized the metrics with the XSUM dataset in mind and checked other available datasets to test the generalizability later. Second, there is a huge methodical difference between the XSUM and SummEval faithfulness annotations. In the XSUM hallucination dataset, annotators worked closely with the text annotating unfaithful passages, whereas in SummEval, annotators used Likert scales, a more distant approach. To exemplify this difference, consider the two sentences "I love you" vs. "I hate you". Using a Likert scale, annotators would most likely rate the summary 1 or 2 (faithfulness score $\leq 25\%$). When using span annotations, the only unfaithful word is "hate", resulting in a faithfulness score of 66%. Both approaches are valid, but for our experiments and quantitative analysis, we stick with the closer, span-annotation-based faithfulness computation.

Method (on XSUM)	Pearson (r)	Spearman (p)
BERTScore	**0.501**	**0.486**
Entailment	0.366	0.422
SentSim	0.392	0.389
SRL	0.393	0.377
NER	0.252	0.259
QGQA	0.228	0.258
Open IE	0.169	0.185
Method (on SummEval)	Pearson (r)	Spearman (p)
SentSim	**0.24**	**0.24**
Entailment	0.22	0.22
BERTScore	0.17	0.17
QGQA	0.13	0.13
SRL	0.13	0.13
NER	0.12	0.12
Open IE	0.10	0.10

Table 2: Pearson (r) and Spearman (p) correlation coefficients for faithfulness measured between human faithfulness judgements and different automatic methods.

Method	Correct	Delta
Random	50.0%	0
NER	29.5%	-20.5
Open IE	49.0%	-1
ESIM (Falke et al., 2019)	67.6%	+17.6
SRL	69.4%	+19.4
SentSim	69.7%	+19.7
FactCC (Kryscinski et al., 2020)	70.0%	+20
QGQA	71.9%	+21.9
BERTScore	77.5%	+27.5
Entailment	**88.5%**	**+38.5**
Human (Falke et al., 2019)	83.9%	+33.9

Table 3: Results on the sentence re-ranking experiment. Human performance was crowd-sourced. Ties are counted as incorrect predictions.

We also evaluate all faithfulness metrics on the sentence re-ranking experiment by Falke et al. (2019). This dataset contains contains 373 triples, each triple consists of a source sentence and two summary sentences. Source sentences are taken from the CNN/DailyMail dataset, summary sentences are generated by the summarization model from Chen and Bansal (2018). One summary sentence is faithful to the source sentence, whereas the other summary sentence is factually inconsistent.

We test how often a metric prefers the correct sentence i.e. gives a higher score to the faithful sentence. Results are shown in Table 3.

Entailment distinguishes best between unfaithful and faithful sentences, achieving 88.5% correct predictions outperforming even human performance. All other faithfulness metrics perform in a comparable range on this task, ranking about 70% example sentences correctly. The only exceptions are Open IE and NER. Both metrics perform worse than Random. We qualitatively find that, in almost every example, the entities mentioned in the summary sentences are also present in the source sentence explaining the poor ranking performance.

Finally, in our search for the best faithfulness metric, we experiment with combining multiple metrics. Since the discussed faithfulness metrics compare fairly different information (tokens, entities, answers to questions etc.), we believe a combination of metrics can lead to a better faithfulness assessment. We correlate all faithfulness metrics with each other using the XSUM hallucination dataset. The results are shown in Figure 1, indicating that a combination of BERTScore, QGQA and either Entailment or NER is promising.

Data to learn a reliable combination of metrics is not available, since manual faithfulness evaluation is time-consuming and expensive. Still, to analyze the effectiveness of combining metrics, we learn a linear combination of multiple metrics with 10-fold cross-validation on the XSUM hallucination dataset. Table 4 shows combining BERTScore, Entailment and QGQA achieves an average Spearman correlation of 0.559, which is a relative improvement of 15% over BERTScore, combining all metrics leads to a relative improvement of 20%.

5 RQ2: Error Analysis of faithfulness metrics

In order to reveal weaknesses and room for improvement, we investigated outputs for 100 randomly selected source-summary pairs of the XSUM hallucination dataset per metric, of which 50 are underprediction cases and 50 are overprediction cases. A detailed breakdown of the most prevalent error categories (E1 - E37) and their relative frequency is shown in Table 5 for all metrics. To set these errors in perspective, Figure 2 visualizes how often, and by how much a metric over- and underpredicts. BERTScore, for example, is much more prone to overpredicting (75%), indicating that these errors are more critical. Next, we discuss ideas to tackle some of the found problems.

The F1 similarity metric is used in many faithfulness metrics (QGQA, SentSim, OpenIE) because it leads to best correlation with human faithfulness.

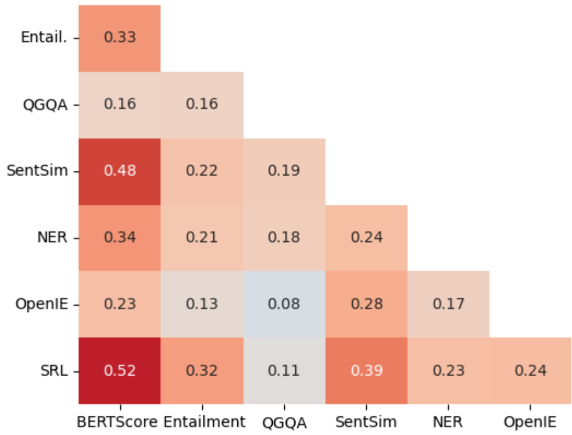

Figure 1: Spearman correlation of faithfulness metrics with each other computed on the XSUM hallucination dataset.

Combination	Correlation
1· BERTScore (BS)	0.485
1.5· BS +0.1· NER	0.493
1.5· BS +0.26· QGQA	0.514
1.3· BS +0.26· Entailment	0.535
1.3· BS +0.24· Entailment +0.24· QGQA	0.559
0.86· BS +0.22· Entailment +0.03· NER +0.21· QGQA + 0.3· SRL +0.34· SS	0.582

Table 4: Averaged Spearman correlations of linear metric combinations with human faithfulness judgements.

This metric performs exact match on a token-level, which comes with many disadvantages: it fails to match synonyms (Error 12 in Table 5), does not comprehend meaning (E14, E29) and stopwords can falsify its results (E24). Further, less frequent errors include inability to correctly compare abbreviations (e.g. "GB" with "Great Britain"), singular and plural (e.g. "men" with "man"), generalizations (e.g. "save 5$" with "save money"),

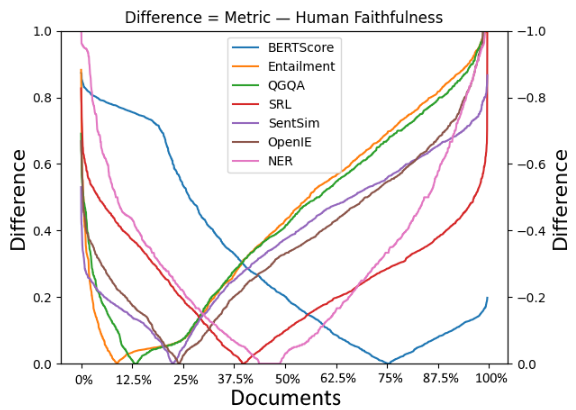

Figure 2: Differences between human and metric faithfulness predictions. Documents and their corresponding difference are sorted in descending order per metric.

locations (e.g. "London" with "England") and e.g. "pharmaceutical firm" with "Accord Healthcare" as it lacks background knowledge. A possible solution is to replace F1 with a metric that has background knowledge and can deal with paraphrases, like BERTScore.

However, the error analysis revealed that BERTScore, which aligns and compares token embeddings, tends to assign too high similarities to phrases that appear in different contexts and to negations, opposites, and contradictions as well as to different numbers. For example, whether someone was jailed for 4 or 7 years makes no difference to BERTScore (similarity of 97%). Currently, BERTScore operates on contextualized embeddings. Paraphrases and synonyms are used in similar context, thus, their embeddings are similar. But, negations, opposites and contradictions typically appear in similar contexts as well, which leads to some of BERTScores problems. Using contrastive embeddings where opposites are distant in the embedding space is a promising direction.

QGQA struggles with questions having not enough variation (E7) or targeting irrelevant information (E9). Questions are generated by providing a model with text and answer candidate, thus, developing an answer candidate selection method that focuses on critical parts of the summary can solve these issues. Further, some generated questions are not answerable, but the QA model finds answers anyway (E8). Here, a QA model that can output "NO ANSWER" is a possible solution.

NER often finds no entities at all (E17) or not enough entities (E20) for the following reason: generated summaries are written in lowercase only. However, one important feature of NER models is capitalization, leading to either not finding entities or incorrect entity labels (E22). Applying a re-capitalization model to generated summaries before extracting entities seems promising.

OpenIE suffers mostly from triples not covering important information (E25). By definition, Open IE triples should cover subject, predicate, object which will always lead to a sentence (or subsentence) representation that misses information. In its current state, we do not think OpenIE is a suitable method to assess faithfulness. Instead, SRL is a solid alternative as these models predict more detailed labels (e.g. who, what, whom, why etc.).

SRL uses cosine similarity of phrase embeddings to align and compare phrases with similar semantics. Similar to BERTScore, cosine similarity of phrases tends to be too high (E30), despite different contexts (E31). We calculate embeddings per phrase and, thus, the remaining sentence has no influence on phrase embeddings. Including more context to the phrase embedding calculations could help issue E31. Other issues attribute to SRL labels. The SRL model predicts wrong labels (E33) or similar summary and source phrases have different labels (E37). We already group SRL labels as described in Section 3.7 to increase robustness and number of matches. Refining this grouping with aid of experts could be beneficial.

The current protocol of SentSim, aligning and comparing one summary with one source sentence, is not a good fit to assess faithfulness (E16). A sophisticated approach that splits sentences into clauses and compares them seems more suitable.

Entailment calculates the entailment probability of a summary sentence given the source document. Analyzing this metric posed quite the challenge as its calculations are in-transparent. We found that verbs have most impact on the predictions: whenever a verb is not entailed, the metric predicts very low scores (E5). Cases where mostly the verbs are unfaithful are problematic as human faithfulness is usually high for summaries that contain few unfaithful words.

6 Conclusion

We re-implemented, modified and proposed new metrics to assess faithfulness of automatically generated summaries. Next, we conducted several experiments and found that BERTscore and Entailment correlate with human judgements and are able to successfully re-rank sentences. In a comprehensive error analysis, we revealed common problems of faithfulness metrics and identified possible solutions to their most prevalent issues. We want to highlight that the discussed metrics do not seem to generalize well to other datasets and cannot replace human faithfulness evaluation yet.

With this work, we laid a solid basis for further development and improvement on faithfulness metrics. We also released an open-source library including all discussed metrics to encourage further experimentation and to facilitate evaluation.

In further work, we experiment with contrastive embeddings and combine multiple metrics to improve performance. Also, we collect new faithfulness datasets to build metrics that generalize well.

#	**BERTScore Errors**	Over	Under
1	Phrases or entities appearing in different context have too high similarity	45%	-
2	Negations, opposites and contradictions have too high similarity	24%	-
3	Different numbers (amounts, counts, money, age, dates etc.) have too high similarity	13%	-
4	Arbitrarily assembled compound nouns have high faithfulness *e.g. "Macedonia's Prime Minister Justin Riot"*	8%	-
#	**Entailment Errors**	Over	Under
5	Faithful phrases connected by unfaithful verbs drastically reduce the score Summary: *Moscow imposed sanctions on Turkey.* Score: *0%* Src: *Russia suspended all sanctions against Turkey.*	-	52%
6	Robustness: summary contains grammatical errors or word repetitions	-	18%
#	**QGQA Errors**	Over	Under
7	Questions do not have enough variation (target the same information, are similar, too few)	44%	48%
8	Question is not answerable, but an answer matching the unfaithful summary is found anyway *Q: Which county has signed Colin? Src: Worcestershire signed John. A: Worcestershire*	32%	-
9	Questions target irrelevant information (answers do not help to assess the faithfulness of the text)	12%	12%
10	QA component cannot find the correct answer	-	36%
11	Question is unanswerable (since no answer can be found, faithfulness decreases)	-	24%
12	F1 answer similarity fails to match correct answers *e.g. "optometrist"* vs. *"eye specialist"* or *"a number of whales"* vs. *"thirty six whales"*	-	44%
#	**SentSim Errors**	Over	Under
13	Stopwords increase the similarity (faithfulness based on stopwords or incorrect alignment)	52%	-
14	F1 does not comprehend meaning (different terms mean the same, or vice versa) *"police appeal for witnesses"* vs. *"anyone with information can call 101"*	14%	36%
15	Summary sentence paraphrases multiple sentences. Comparing with one sentence is insufficient.	32%	56%
16	Erroneous sentence splitting (information is wrongly split into multiple sentences)	-	12%
#	**NER Errors**	Over	Under
17	No entities in the summary (faithfulness defaults to 100%)	50%	-
18	No source entities with corresponding tag to summary entity ($\rightarrow$ not considered in calculation)	16%	-
19	Entities match correctly, but faithfulness is not related to entities	14%	30%
20	Important entities not found in summary and / or source (*e.g. Leukaemia not detected as entity*)	26%	61%
21	Tokenization problems lead to incorrect entities (*e.g. 1.5million = 1[Money].5m[Quantity]*)	-	12%
22	Incorrect entity labels (*e.g. World is labeled as Person*)	-	12%
23	Similarity of different mentions of same entity is low (*e.g. "Myles Anderson"* vs. *"Anderson"*)	-	24%
#	**OpenIE Error**	Over	Under
24	Stopwords increase the similarity of completely different triples	40%	-
25	Summary triples miss important information (dates, locations, etc.) *e.g. a man* | *has been* | *found* instead of *a man* | *has been found guilty* | *of murdering a soldier* *"More than a third of children in the UK have been sexually abused"* $\rightarrow$ *Children* | *in* | *UK*	44%	52%
26	Faithful information of source document not part of a triple	-	26%
27	Summary is too abstract (highly paraphrased, aggregate information of multiple sentences)	-	20%
28	Summary has no triples	-	16%
29	F1 does not comprehend meaning (different terms mean the same, or vice versa)	-	8%
#	**SRL Errors**	Over	Under
30	Similarity of (apparently randomly) aligned phrases is incomprehensibly high	44%	-
31	Single word phrases match exactly with other single word phrases, but context is different	28%	-
32	Similarity of detailed, information-rich summary phrases and simple source phrases is too high *e.g. "Double olympic champion Nicola Adams"* is very similar to *"Adams"*	16%	-
33	SRL model errors (incorrect labels, incorrect split of phrases, incorrect grouping of phrases) *e.g. "IS"* (abbreviation of islamic state) or *"united"* of *"Manchester United"* is labeled as verb	12%	-
34	Important information is not part of a phrase and cannot be considered in faithfulness calculation	16%	-
35	Summary phrases are coarse grained. Split into smaller phrases necessary to validate faithfulness	-	40%
36	Summary is too abstract (understanding of whole text necessary to validate faithfulness) *e.g.* summary presents the result of a soccer match, source is soccer live ticker	-	24%
37	Faithful phrases have different tags in summary & source and, thus, are not aligned & compared	-	32%

Table 5: Quantitative error analysis of 100 randomly selected examples of the XSUM hallucination dataset for all faithfulness metrics, of which 50 are underprediction (Under) and 50 are overprediction (Over) cases.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, pages 4784–4791, New Orleans, Louisiana, USA.

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.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, page 177–190.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. 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 Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. 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 International Conference on Knowledge Discovery + Data Mining*, page 166–175, New York, New York, USA.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, page 1693–1701.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

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.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Computation and Language repository, arXiv:1907.11692.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. 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.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. Computation and Language repository, arXiv:1904.05255.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339, Vienna, Austria.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *Proceedings of the 8th International Conference on Learning Representations*, Accepted as poster. Online.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020b. 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.

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

A Appendix

A.1 Comparing texts

Most faithfulness metrics introduced in Section 3 compare texts to compute the faithfulness score. We experiment with various similarity metrics to implement the faithfulness metrics and evaluate them on the XSUM hallucination dataset (Table 7 and the sentence re-ranking experiment (Table 8). The cosine-similarity (CS) metric is calculated on sentence embeddings generated by off-the-shelf sentence-transformers[5]. We find using F1 in QGQA is the best trade-off between performance and computation time. SRL performs best with CS. Depending on the task, NER performs best with either F1 or CS. Both, SentSim and Open IE perform best with either F1 or BERTScore.

A.2 Input for textual entailment

We evaluate different input techniques (sentence-to-sentences (s2s), document-to-sentence(d2s), top-to-sentence (top2s) for an entailment model on the XSUM hallucination dataset and find that d2s works best as shown in Table 6.

Method	Pearson (r)	Spearman (p)
s2s	0.152	0.190
d2s	**0.366**	**0.422**
top2s	0.251	0.302

Table 6: Evaluation of different input techniques for entailment models. The table lists correlations with human faithfulness judgements.

Method	Similarity	Pearson (r)	Spearman (p)
QGQA	EM	0.200	0.226
QGQA	F1	0.228	0.258
QGQA	**BERTScore**	**0.252**	**0.258**
QGQA	CS	0.216	0.222
NER	EM	0.251	0.255
NER	**F1**	**0.252**	**0.259**
NER	BERTScore	0.151	0.195
NER	CS	0.200	0.204
SRL	EM	0.234	0.273
SRL	F1	0.359	0.363
SRL	BERTScore	0.270	0.344
SRL	**CS**	**0.393**	**0.377**
SentSim	EM	-0.039	-0.039
SentSim	**F1**	**0.392**	**0.389**
SentSim	BERTScore	0.374	0.372
SentSim	CS	0.387	0.369
Open IE	EM	0.042	0.076
Open IE	**F1**	**0.169**	**0.185**
Open IE	BERTScore	0.013	0.212
Open IE	CS	0.134	0.186

Table 7: Comparison of different similarity metrics used in various faithfulness metrics. The table lists correlations with human faithfulness judgements. We experiment with Exact Match (EM), F1 (on token-level), BERTScore and cosine-similarity of embeddings (CS).

Method	**Similarity**	**Correct**
QGQA	EM	67.29%
QGQA	F1	68.36%
QGQA	BERTScore	69.17%
QGQA	**CS**	**69.71%**
NER	EM	18.50%
NER	F1	18.50%
NER	BERTScore	26.54%
NER	**CS**	**29.49%**
SRL	EM	50.67%
SRL	F1	66.76%
SRL	BERTScore	67.83%
SRL	**CS**	**69.44%**
SentSim	EM	2.95%
SentSim	F1	56.03%
SentSim	**BERTScore**	**69.71%**
SentSim	CS	68.36%
Open IE	EM	26.27%
Open IE	F1	46.11%
Open IE	**BERTScore**	**49.06%**
Open IE	CS	47.99%
Open IE	RMR1	21.98%
Open IE	RMR2	26.27%

Table 8: Comparison of different similarity metrics used in various faithfulness metrics evaluated on the sentence ranking experiment from Falke et al. (2019). We experiment with Exact Match (EM), F1 (on token-level), BERTScore and cosine-similarity of embeddings (CS).

[5] https://www.sbert.net/index.html