Understanding the Extent to which Content Quality Metrics Measure the Information Quality of Summaries

Daniel Deutsch and Dan Roth
Department of Computer and Information Science
University of Pennsylvania
ddeutsch,danroth@seas.upenn.edu

Abstract

Reference-based metrics such as ROUGE or BERTScore evaluate the content quality of a summary by comparing the summary to a reference. Ideally, this comparison should measure the summary's information quality by calculating how much information the summaries have in common. In this work, we analyze the token alignments used by ROUGE and BERTScore to compare summaries and argue that their scores largely cannot be interpreted as measuring information overlap. Rather, they are better estimates of the extent to which the summaries discuss the same topics. Further, we provide evidence that this result holds true for many other summarization evaluation metrics. The consequence of this result is that the most frequently used summarization evaluation metrics do not align with the community's research goal, to generate summaries with high-quality information. However, we conclude by demonstrating that a recently proposed metric, QAEval, which scores summaries using question-answering, appears to better capture information quality than current evaluations, highlighting a direction for future research.[1]

1 Introduction

The development of a reliable metric that can automatically evaluate the content of a summary has been an active area of research for nearly two decades. Over the years, many different metrics have been proposed (Lin, 2004; Hovy et al., 2006; Giannakopoulos et al., 2008; Louis and Nenkova, 2013; Zhao et al., 2019; Zhang et al., 2020; Deutsch et al., 2021).

The majority of these of approaches score a candidate summary by measuring its similarity to a reference summary. The most popular metric, ROUGE (Lin, 2004), compares summaries based on their lexical overlap, whereas more recent methods, such as BERTScore (Zhang et al., 2020), do so based on the similarity of the summary tokens' contextualized word embeddings.

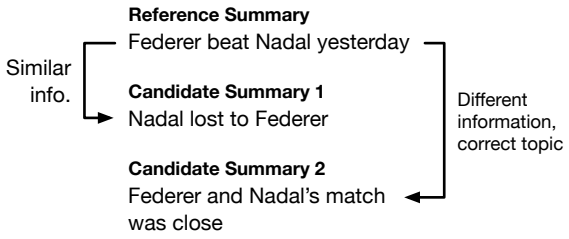


Figure 1: Both candidate summaries are similar to the reference, but along different dimensions: Candidate 1 contains some of the same information, whereas candidate 2's information is different, but it at least discusses the correct topic. The goal of this work is to understand if summarization evaluation metrics' scores should be interpreted as measures of information overlap or, less desirably, topic similarity.

Ideally, metrics that evaluate the content of a summary should measure the quality of the information in the summary. However, it is not clear whether metrics such as ROUGE and BERTScore evaluate summaries based on how much information they have in common with the reference or some less desirable dimension of similarity, such as whether the two summaries discuss the same topics (see Fig. 1).

In this work, we demonstrate that ROUGE and BERTScore largely do not measure how much information two summaries have in common. Further, we provide evidence that suggests this result holds true for many other evaluation metrics as well. Together, this allows us to conclude that the most frequently used metrics in summarization fail to evaluate the quality of information in a summary.

Our analysis casts ROUGE and BERTScore into a unified framework in which the similarity of two summaries is calculated based on an alignment be-

[1]Our code is available at https://github.com/CogComp/content-analysis-experiments.

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 300–309
November 10–11, 2021.

tween the summaries' tokens (§3). This alignment-based view of the metrics enables performing two different analyses of how well they measure information overlap.

The first analysis demonstrates that only a small proportion of the metrics' token alignments are between phrases which contain identical information according to domain experts (§4). The second reveals that token alignments which represent common information are vastly outnumbered by those which represent the summaries discussing the same topic (§5). Overall, both analyses support the conclusion that ROUGE and BERTScore largely do not measure information overlap.

Then, we expand our analysis to consider if 10 other evaluation metrics successfully measure information quality or not (§6). By demonstrating that nearly all of the metrics correlate much more strongly to ROUGE than to the gold-standard method of manually comparing two summaries' information (Nenkova and Passonneau, 2004), we argue the metrics are likely to measure information overlap no better than ROUGE does.

However, one recently proposed metric that evaluates summaries using question-answering (QA), QAEval (Deutsch et al., 2021), correlates equally well to ROUGE and gold-standard annotations of information overlap. By viewing QAEval as inducing an alignment between two summaries and reasoning about its behavior, we demonstrate evidence that it measures information overlap much more strongly than either ROUGE or BERTScore does, supporting that QA-based metrics are a promising direction for future research (§7).

While the summarization community has been aware, informally, of the shortcomings of the current evaluation metrics, this study provides experimental evidence beyond correlations to support these intuitions. The contributions of this work include (1) an analysis which reveals that ROUGE and BERTScore largely do not measure the information overlap between two summaries, (2) evidence that many other evaluation metrics likely suffer from the problem, and (3) preliminary results which show that QAEval does measure information quality better than ROUGE and BERTScore.

2 Understanding Evaluation Metrics

Reference-based evaluation metrics assume that human-written reference summaries have gold-standard content and score a candidate summary based on its similarity to the reference. An ideal evaluation metric that measures the content quality of a summary should score the quality of its *information*. For reference-based metrics, this means that the comparison between the two summaries should measure how much information they have in common.

In this work, our definition of information is equivalent to what can be expressed through predicate-argument relations in text. Information can be expressed in other ways, such as entailment, but we focus on predicates and arguments since they are more easily directly evaluated.

Metrics such as ROUGE and BERTScore calculate the similarity of two summaries either by how much lexical overlap they have or how similar the summaries' contextual word embeddings are (discussed in more detail in §3). Although we understand how their scores are calculated, it is not clear how the scores should be interpreted: Are they representative of how much information the two summaries have in common, or do they describe how similar the summaries are on some other less desirable dimension, such as whether they discuss the same topics? Our goal is to answer this question.

Knowing the answer is critically important. The goal of summarization is to produce summaries which contain the "correct" information (among other desiderata). Automatic metrics are the most frequent method that researchers use to argue that one summarization model generates better summaries than another. If our evaluation metrics are not aligned with our research goals — or if we do not understand what they measure at all — then we do not know whether we are making progress as a community.

3 A Common Framework

The focus of our analysis will be primarily on two evaluation metrics, ROUGE and BERTScore. Although on the surface these two metrics appear to compare two summaries very differently, here we demonstrate how they can both be viewed as calculating a score based on a weighted alignment between the summaries' tokens. This common framework enables us to reason about how to interpret their scores.

Let $R = r_1, r_2, \ldots, r_m$ and $S = s_1, s_2, \ldots, s_n$ be the tokens of the reference and candidate summaries. ROUGE-1 counts the number of unigrams

that are in common between the two summaries:[2]

$$M = \sum_{s \in S} \min(c_R(s), c_S(s)) \quad (1)$$

where $c_T(s)$ counts the number of times s appears in the summary T and the summand is over unique unigrams. Then, precision and recall are calculated by dividing M by n and m, respectively. When multiple references are available, the precision and recall scores are micro-averaged.

A weighted alignment A is a set of token alignments (i, j, w) that map token r_i to s_j with weight $w \in (0, 1.0]$. The weight of an alignment, denoted $W(A)$, is the sum of the weights of the individual token alignments. ROUGE can be viewed as creating an alignment by pairing $\min(c_R(s), c_S(s))$ occurrences of unigram s in R and S with weight 1.0 for all unigrams. It additionally imposes a constraint that each token can be aligned to at most one other token. Since a unigram may appear multiple times in a summary, the alignment may not be unique, however, its weight will equal M.

BERTScore calculates a similarity score between two pieces of text based on the pairwise cosine similarities of their tokens' BERT embeddings. Let B_{ij} be the similarity score between the embeddings for r_i and s_j. To calculate recall, BERTScore first aligns every reference token to its most-similar summary token (Eq. 2). Then, the sum of the corresponding similarities is normalized by the number of reference tokens to get the recall score (Eq. 3).

$$A_R = \{(i, j, B_{ij}) : \forall i, j = \arg\max_k B_{ik}\} \quad (2)$$

$$\text{BERTScore}_{\text{Recall}} = W(A_R)/m \quad (3)$$

A similar procedure is followed to calculate precision, but instead, every summary token is aligned to its most-similar reference token, and the sum of the similarities is normalized by the number of summary tokens. When multiple references are available, the precision and recall scores are defined to be the maximum respective values across references. Because the token alignments are selected via the max operation, BERTScore's alignment is unique, unlike for ROUGE.

By formulating ROUGE and BERTScore in a framework based on token alignments, we can reason about their behaviors by examining the tokens they align in two different analyses, as described next.

[2]Our analysis focuses on the unigram variant of ROUGE, called ROUGE-1. We refer to it, where clear, as ROUGE for simplicity.

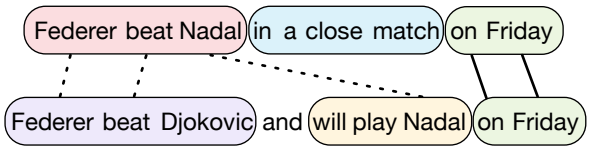


Figure 2: An example token alignment used by ROUGE or BERTScore. Each color represents a summary content unit (SCU) that marks informational content. Only $2/5$ of the token alignments (the solid edges) can be explained by matches between phrases that express the same information (the green phrases).

4 SCU-Based Analysis

The first analysis compares the two metrics' token alignments to annotations derived from the Pyramid Method (Nenkova and Passonneau, 2004).

The Pyramid Method is a technique to manually evaluate the content of a candidate summary by comparing it to a set of reference summaries. The method uses a domain-expert annotator to exhaustively identify atomic units of meaning in the summaries, known as summary content units (SCUs), and mark their occurrences in the reference and candidate summaries. Two phrases marked with the same SCU are considered to express the same information. Since the Pyramid Method annotation is exhaustive, we can assume that any two phrases in the reference and candidate summaries that are not marked with the same SCU do not have the same meaning.

These annotated phrases can be used to reason about ROUGE and BERTScore: If a large proportion of their token alignments is between phrases that express the same information, then their scores can potentially be interpreted as representing the summaries' information overlap. Otherwise, it is evidence that they do not compare summaries based on information.

For this analysis, we use the summaries and Pyramid annotations from the TAC 2008 and 2009 English multi-document summarization datasets (Dang and Owczarzak, 2008, 2009). TAC 2008 has 48 document clusters and 58 system summaries, and TAC 2009 has 44 clusters and summaries from 55 systems. All clusters have around 10 documents each and 4 reference summaries, and every summary has been been annotated with SCUs.

For each of the system summaries, we calculate the proportion of the total alignment weight

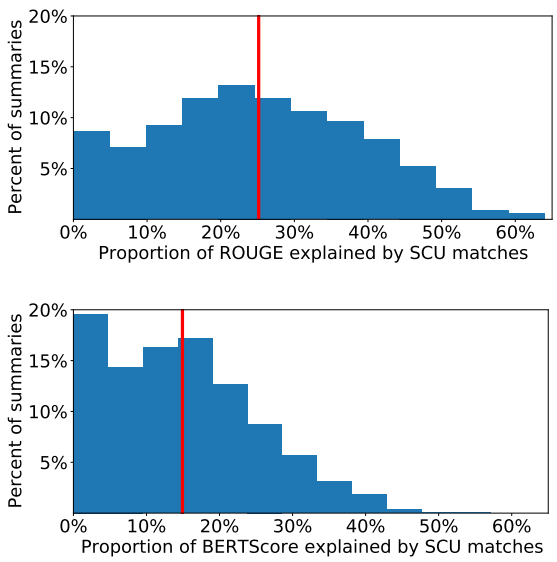


Figure 3: The distribution of the proportion of ROUGE (top) and BERTScore (bottom) on TAC 2008 that can be explained by tokens matches that are labeled with the same SCU (Eq. 5). The averages, 25% and 15% (in red), indicate that only a small amount of their scores is between phrases that express the same information.

that can be explained by matches between identical SCUs, as defined in Eqs. 4 and 5:

$$A_{\text{SCU}} = \{(i,j,w) : (i,j,w) \in A, \\ \text{SCU}(i) \cap \text{SCU}(j) \neq \emptyset\} \quad (4)$$

$$\text{Prop}_{\text{SCU}} = \frac{W(A_{\text{SCU}})}{W(A)} \quad (5)$$

where $\text{SCU}(i)$ returns the set of SCUs that are annotated for the token at index i. Fig. 2 has an example of this calculation. Since ROUGE does not use a unique alignment, we choose the alignment which maximizes Eq. 5, thus calculating an upper-bound.

The distribution of the proportion of ROUGE and BERTScore explained by SCU matches is presented in Figure 3. We find that, on average, only 25% and 15% of these metrics scores comes from matches between tokens marked with the same SCUs. Since only a relatively small fraction of the overall metric scores comes from phrases with the same information, this suggests that ROUGE and BERTScore's values cannot be interpreted as a measure of information overlap.

5 Category-Based Analysis

The second analysis of ROUGE and BERTScore focuses on grouping token alignments into categories (§5.1), then using those categories to reason about how much of the metrics' scores is explained by information or topic matches (§5.2).

5.1 Token Alignment Categorization

We define a category to be a function C that selects the subset of summary token indices for which that category applies. For example, a "noun" category would select only the token indices that correspond to nouns. $C(S)$ denotes the application of a category to summary S.

Each category is used to filter an alignment A used by ROUGE or BERTScore to a category-specific alignment between tokens which belong to that category only, denoted A_C:

$$A_C = \{(i,j,w) : (i,j,w) \in A, \\ i \in C(R), j \in C(S)\} \quad (6)$$

For the "noun" category, A_C would be the subset of token alignments between nouns in R and S.

Then, the *contribution* of C is defined as the ratio between A_C and A:

$$\text{Contribution}_C = \frac{W(A_C)}{W(A)} \quad (7)$$

The contribution of C can be interpreted as the portion of ROUGE or BERTScore that can be explained by matches between tokens in category C (see Fig. 4 for an example).

Higher-Order Categories Although our analysis only uses unigram alignments, it is desirable to reason about groups of tokens. This would enable calculating how much of the metrics' scores can be explained by matches between (subject, verb, object) tuples, for instance.

We extend the definition of a category to select a set of *tuples* of indices. Then A_C selects only the token alignments in A that are included in an aligned tuple selected by C. Two tuples $(i_1, \ldots, i_k)$ and $(j_1, \ldots, j_k)$ are said to be aligned if indices i_ℓ and j_ℓ are aligned for $\ell = 1, \ldots, k$. Fig. 5 has an example tuple-based matching.

5.2 Category-Based Analysis

Next, we define a set of categories in which each category represents either information or topic matches, then reason about how much of the metrics' scores can be explained by information or topic similarities based on the corresponding category contributions.

We define the following categories:

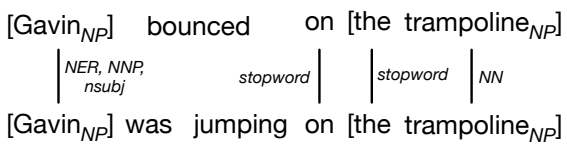


Figure 4: Every token alignment used by ROUGE or BERTScore is assigned to one or more interpretable categories (defined in §5). This allows us to calculate, for this example, that matches between named-entities contribute $1/4$ to the overall score, stopwords $2/4$, and noun phrases $3/4$ (assuming alignment weights of 1.0).

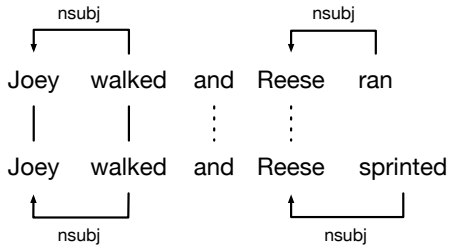


Figure 5: The VB+NSUBJ category selects tuples of verbs and their corresponding NSUBJ dependents in the dependency tree. In this example, $2/4$ of the alignment (the solid lines) can be explained by matches between such tuples. The dashed lines cannot: The "and" alignment is not part of any tuple; Since "ran" and "sprinted" are not aligned, their corresponding tuples are not considered to be aligned, so the "Reese" match does not count toward the total.

1. **Stopwords**: One category to select matches between stopwords, denoted STOPWORDS.

2. **Parts-of-Speech**: Six categories, one for selecting alignments between each type of the following part-of-speech tags: common nouns (NN), proper nouns (NNP), verbs (VB), adjectives (ADJ), adverbs (ADV), and numerals (NUM).

3. **Named-Entity**: One category for all named-entities, denoted NER. This category only selects alignments between tokens if they are the same type of named-entity (person, location, or organization).

4. **NP Chunks**: One category to select matches between tokens that are part of noun phrases, denoted NP-CHUNKS.

5. **Dependency**: Three categories that select matches between tokens with the same dependency tree arc label for ROOT, NSUBJ, and DOBJ labels.

	TAC'08		CNN/DM	
Category	**R**	**BS**	**R**	**BS**
NP-CHUNKS	58.7	46.1	53.6	43.0
STOPWORDS	54.6	32.4	48.4	28.7
NN	17.9	13.7	31.8	24.9
NNP	14.9	11.3	0.3	0.2
NER	13.5	8.5	0.1	0.1
VB	9.0	9.3	14.1	10.6
ADJ	4.1	2.6	6.2	4.0
NSUBJ	3.9	2.2	6.3	4.1
DOBJ	2.0	1.4	2.8	1.7
NUM	1.5	1.7	2.5	1.9
VB+DOBJ	1.3	0.4	3.4	1.0
ROOT	1.1	1.5	3.3	2.5
VB+NSUBJ	1.0	0.5	3.8	2.4
ADV	0.6	0.4	1.6	0.8
VB+NSUBJ+DOBJ	0.3	0.1	1.5	0.5

Table 1: The contributions (Eq. 7) of every category to ROUGE (R) and BERTScore (BS) on TAC 2008 and CNN/DailyMail indicate the metrics are largely matching nouns and stopwords rather than tuples which express information (e.g., VB+NSUBJ+DOBJ). The contributions do not sum to 100% because more than one category can explain the same token alignment. The NNP and NER for CNN/DailyMail are significantly lower because the candidate summaries were all lower-cased.

6. **Dependency Tuples**: Three categories that match higher-order tuples based on the dependency tree. Each category selects a tuple containing a verb and either its subject child (NSUBJ), object child (DOBJ), or both. These categories are denoted VB+NSUBJ, VB+DOBJ, and VB+NSUBJ+DOBJ. They are representative of information expressed as predicate-argument relations (e.g., {subject, verb, object} tuples).

We consider 2 through 5 to be keywords that represent the topics discussed in the summaries, whereas 6 describes tuples which express the summaries' information as predicate-argument relations.

The contributions of each category on the TAC 2008 summaries as well as the summaries produced by baseline (See et al., 2017) and state-of-the-art (Liu and Lapata, 2019) abstractive models on the CNN/DailyMail dataset (Nallapati et al., 2016) is presented in Table 1. The POS/NER tagging and parsing are all done with spaCy (Honnibal et al., 2020).

The results across datasets and evaluation met-

Content Type	TAC'08 R	TAC'08 BS	CNN/DM R	CNN/DM BS
Topic	70.6	57.9	75.0	59.2
Information	2.2	0.9	6.7	3.2
Stopwords	54.6	32.4	48.4	28.7

Table 2: The contributions of different categories of token matches when grouped by whether they represent topics, information, or stopwords. Clearly, the information categories explain only a small proportion of the overall metrics scores on TAC'08 and CNN/DailyMail.

rics largely follow the same trend: Noun- and stopword-based matches explain the vast majority of the token alignments used by both ROUGE and BERTScore, whereas the dependency tuple categories explain very little of the overall scores.[3] For instance, on TAC 2008, noun phrase and stopword matches contribute 58.7% and 54.6% to ROUGE, whereas the dependency tuple with the largest contribution, VB+DOBJ only contributes 1.3%.

When the specific categories are grouped by content type in Table 2, it becomes even more apparent that topic and stopwords matches explain most of ROUGE and BERTScore.[4] We find that topic, stopword, and information matches explain 70.6%, 54.6%, and 2.2% of ROUGE on TAC'08.

The low contribution of information-based categories toward each metric is further evidence that neither metric strongly captures the information overlap between summaries, supporting the results found in §4.

6 Other Evaluation Metrics

The analyses thus far have exploited the structure of ROUGE and BERTScore to reason about the extent to which they measure information overlap between two summaries. Although it is desirable to ask the same question about other evaluation metrics, the metrics may not directly fit into this analysis framework or it would require significant effort to repeat this analysis for each one. Instead, we indirectly reason about how much information overlap other metrics measure through their correlations to ROUGE and the Pyramid Score.

[3] Although there are versions of ROUGE that remove stopwords, including them is significantly more common, and therefore we analyze the more popular ROUGE variant.

[4] The numbers in Table 2 numbers cannot be directly read off Table 1 nor do they sum to 100% because multiple categories can explain the same token alignment.

Metric	ROUGE-1	Pyr. Score	Δ
ROUGE-1	1.00	0.59	-
Pyramid Score	0.59	1.00	-
AutoSummENG	0.83	0.61	0.22
BERTScore	0.74	0.59	0.15
BEwT-E	0.81	0.62	0.19
MeMoG	0.68	0.52	0.16
METEOR	0.91	0.63	0.28
MoverScore	0.79	0.61	0.18
NPowER	0.81	0.60	0.21
PyrEval	0.47	0.35	0.12
QAEval-F_1	0.59	0.57	0.02
ROUGE-2	0.79	0.58	0.21
S^3	0.92	0.63	0.29

Table 3: The summary-level Pearson correlations of various metrics to ROUGE-1 and the Pyramid Score (Δ is the difference between them). All of the other metrics correlate more strongly to ROUGE-1 than the Pyramid Score (by ≈0.2) and correlate to the Pyramid Score approximately as much as ROUGE-1 does (≈0.6). Together, these results suggest the other metrics measure information overlap as poorly as ROUGE-1.

First, we assume that the Pyramid Score is the gold-standard for measuring the information overlap between summaries. This is a relatively safe assumption because the Pyramid Method is annotated by domain experts, and a candidate's Pyramid Score is based solely on how much information it has in common with a reference. There is no credit given to a candidate for discussing the right topics but with the incorrect information.

Then, the correlations of the other metrics to both ROUGE and the Pyramid Score are calculated and compared. If the correlation to ROUGE is much higher than the correlation to the Pyramid Score, then it is more likely that the metric suffers from the same issues that ROUGE does than it is to directly measure information overlap.

Table 3 contains the summary-level correlations of various other evaluation metrics to ROUGE and the Pyramid Score. The other metrics are: AutoSummENG (Giannakopoulos et al., 2008), BEwT-E (Tratz and Hovy, 2008), MeMoG (Giannakopoulos and Karkaletsis, 2011), METEOR (Denkowski and Lavie, 2014), MoverScore (Zhao et al., 2019), NPowER (Giannakopoulos and Karkaletsis, 2013), PyrEval (Gao et al., 2019), QAEval (Deutsch et al., 2021), ROUGE-2, and S^3 (Peyrard et al., 2017). These metrics exhibit a variety of different comparison techniques, from n-gram graph compar-

isons to contextual word-embedding comparisons, other alignment based approaches, and question-answering.

Notably, most of the metrics' Pearson correlations to ROUGE are much higher than to the Pyramid Score by around 0.2 points, suggesting these metrics do not measure information overlap well. Further, their correlations to the Pyramid Score are roughly the same as ROUGE's, around 0.6. This means that these metrics correlate to a direct measure of information overlap as well as one would expect a metric which measures information overlap at the level of ROUGE to correlate. Although the results of this experiment are not direct evidence that many of the other evaluation metrics do a poor job at measuring information overlap, they do strongly suggest it.

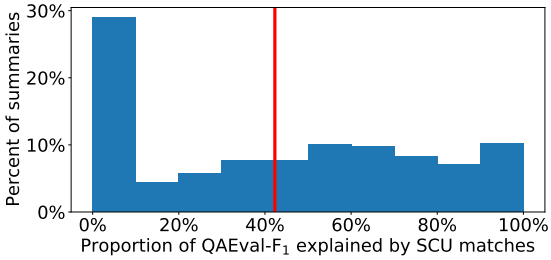
One metric, however, does appear to behave differently from the others. The difference between the correlations for QAEval-F_1 is only 0.02, the smallest among the metrics. In the next section, we run a preliminary analysis of QAEval to understand if this recently proposed metric behaves different than ROUGE and BERTScore.

7 Examining QAEval

Among the metrics analyzed in §6, QAEval stands out not only because it is nearly as similar to ROUGE as it is to the Pyramid Score, but also because its approach, evaluating summaries using QA, is rather different than the other methods, which largely rely on measuring textual overlap. In this section, we briefly examine QAEval and demonstrate evidence that it addresses some of the shortcomings of ROUGE and BERTScore.

QAEval is a reference-based evaluation metric proposed by Deutsch et al. (2021) that uses QA to evaluate a summary. The metric automatically generates a wh-question for every noun phrase in the reference summary, then predicts an answer span in the candidate summary for each question if the QA model predicts an answer exists. The score of the metric is the proportion of those questions which are answered correctly, as evaluated by the SQuAD F_1 string similarity (Rajpurkar et al., 2016).

The mapping between the noun phrase in the reference summary and the predicted span in the candidate summary for the corresponding question facilitates viewing QAEval as inducing an alignment between the two summaries. The final metric score is the total weight of the alignment, where the weight of an edge is the F_1 score between the two spans, normalized by the number of questions. Since QAEval can be viewed as a weighted alignment, we are able to repeat the analysis from §4 for QAEval and measure what proportion of its score can be explained by matches between SCUs.[5]



Metric	Proportion of Metric			
	0-25%	25-50%	50-75%	75-100%
ROUGE	51%	45%	4%	0%
BERTScore	83%	17%	0%	0%
QAEval	36%	18%	24%	22%

Figure 6: (Top) The distribution of the proportion of the QAEval-F_1 score that is explained by SCU matches. (Bottom) The percentage of summaries with a score explained by a given proportion of SCU matches. We find that QAEval can be explained by SCU matches far more than ROUGE or BERTScore on average.

The distribution of the proportion of QAEval-F_1 that can be explained by SCU matches on TAC 2008 is shown in Fig. 6. The average proportion, 42%, is higher than ROUGE (25%) and BERTScore (15%), indicating QAEval captures information similarity better than the other two metrics. The table in Fig. 6 summarizes this distribution for QAEval, ROUGE, and BERTScore (Fig. 3), demonstrating that 46% of summaries have at least 50% of their QAEval score explained by SCU matches, whereas this is true for less than 4% and 0% of ROUGE and BERTScore.

Overall, this experiment provides preliminary evidence that QAEval is indeed measuring information quality more than either ROUGE or BERTScore, and suggests that QA-based evaluation metrics are an exciting direction for future research.

[5]We do not repeat the category-based analysis from §5 because QAEval induces a mapping between noun phrases, so analyzing information-based categories, which include predicates, would not produce an interesting result.

Metric	Summ-Level		Sys-Level	
	r	ρ	r	ρ
ROUGE	0.49	0.48	0.80	0.80
NP-CHUNKS	0.45	0.44	0.79	0.80

Table 4: The Pearson r and Spearman ρ correlations of ROUGE and calculating ROUGE with only NP chunks are very close, demonstrating that a purely topic based comparison (NP chunks) is a very high baseline for content quality correlations on TAC'08.

8 Discussion

Responsiveness Correlations Many of the automatic metrics analyzed in this work have demonstrated very high system-level correlations to ground-truth summary responsiveness judgments (Pearson's $r > 0.8$; Dang and Owczarzak, 2008, 2009), so the results that indicate they do not measure information overlap are somewhat contradictory. Since the metrics appear to compare summaries based on the topics they discuss, it is likely that only comparing summary topics is a very strong baseline for these benchmark datasets.

Indeed, we find in Table 4 that calculating ROUGE with only NP chunks (which represents little-to-no information) achieves nearly the same correlations as ROUGE on TAC'08. It is clear that this is not a good evaluation metric, but it does demonstrate that the baseline for this task is quite high.

Limitations There are some limitations to our analysis. First, the results are specific to the datasets and summarization models that were used. However, TAC'08 and '09 are the benchmark datasets for evaluating content quality and have been widely used to measure the performance of different metrics. Further, because the results from §5.2 are consistent across two rather different datasets (TAC and CNN/DailyMail), we believe these results are likely to hold for other datasets.

Then, the predicate-argument based information categories from §5.2 do not capture all of the information from a summary. A phrase like "the Turkish journalist" expresses the nationality of the journalist, but this information would not be represented by the tuples included in our analysis. However, we do not believe the addition of more tuples that express information outside of predicate-argument relationships would significantly change the experimental results.

9 Related Work

Most of the work that reasons about how to interpret the scores of evaluation metrics does so indirectly through correlations to human judgments (Dang and Owczarzak, 2009; Owczarzak and Dang, 2011). However, a high correlation is not conclusive evidence about what a metric measures since it is possible for the metric to directly measure some other aspect of a summary, which is in turn correlated with the ground-truth judgments (see §8). Our work can be viewed as more direct evidence about what ROUGE and BERTScore measure.

Recent work by Wang et al. (2020) argues that many of the same evaluation metrics covered in this work do not successfully measure the faithfulness of a summary based on low correlations to ground-truth judgments. The results from our experiments offer an explanation for why this is the case: The metrics do not compare summaries based on their information, therefore they cannot determine if a summary is factually consistent with its input.

Metrics which do attempt to directly measure information overlap between summaries are based on the gold-standard comparison technique, the Pyramid Method (Nenkova and Passonneau, 2004). Although it relies heavily on annotations by experts, there have been attempts to crowdsource (Shapira et al., 2019) or automate all or parts of the Pyramid Method (Passonneau et al., 2013; Yang et al., 2016; Hirao et al., 2018) including PyrEval (Gao et al., 2019), which we analyzed in §6. These metrics have been met with less success than the text overlap-based ones covered by this work, potentially because measuring information overlap is more difficult than comparing summaries by their topics, and topic-based evaluations strongly correlate to responsiveness judgments (see §8).

10 Conclusion

In this work, we argued that ROUGE, BERTScore, and many other proposed metrics for evaluating the content quality of summaries largely do not compare summaries based on their information overlap. The implications of this result are that the summarization community does not have a reliable metric that aligns with its research goal, to generate summaries with high-quality information. However, we demonstrate that QAEval does appear to measure information overlap better than text overlap-based metrics, highlighting QA-based evaluations as a promising direction for future research.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback on our work.

This work was partly supported by a Focused Award from Google and by contracts FA8750-19-2-1004 and FA8750-19-2-0201 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. In *Proceedings of the First Text Analysis Conference, TAC 2008, Gaithersburg, Maryland, USA, November 17-19, 2008*. NIST.

Hoa Trang Dang and Karolina Owczarzak. 2009. Overview of the TAC 2009 Summarization Track. In *Proceedings of the Text Analysis Conference*.

Michael J. Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 376–380. The Association for Computer Linguistics.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019. Automated Pyramid Summarization Evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 404–418. Association for Computational Linguistics.

George Giannakopoulos and Vangelis Karkaletsis. 2011. Autosummeng and memog in evaluating guided summaries. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*. NIST.

George Giannakopoulos and Vangelis Karkaletsis. 2013. Summary evaluation: Together we stand npower-ed. In *Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*, volume 7817 of *Lecture Notes in Computer Science*, pages 436–450. Springer.

George Giannakopoulos, Vangelis Karkaletsis, George A. Vouros, and Panagiotis Stamatopoulos. 2008. Summarization System Evaluation Revisited: N-Gram Graphs. *ACM Trans. Speech Lang. Process.*, 5(3):5:1–5:39.

Tsutomu Hirao, Hidetaka Kamigaito, and Masaaki Nagata. 2018. Automatic pyramid evaluation exploiting edu-based extractive reference summaries. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4177–4186. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Eduard H. Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated Summarization Evaluation with Basic Elements. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*, pages 899–902. European Language Resources Association (ELRA).

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3728–3738. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):267–300.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 145–152. The Association for Computational Linguistics.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the TAC 2011 summarization track: Guided task and AESOP task. In *Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November*.

Rebecca J. Passonneau, Emily Chen, Weiwei Guo, and Dolores Perin. 2013. Automated pyramid scoring of summaries using distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 143–147. The Association for Computer Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to Score System Summaries for Better Content Selection Evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 74–84. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing Lightweight Pyramids for Manual Summary Evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 682–687. Association for Computational Linguistics.

Stephen Tratz and Eduard H. Hovy. 2008. Summarization Evaluation Using Transformed Basic Elements. In *Proceedings of the First Text Analysis Conference, TAC 2008, Gaithersburg, Maryland, USA, November 17-19, 2008*. NIST.

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, ACL 2020, Online, July 5-10, 2020*, pages 5008–5020. Association for Computational Linguistics.

Qian Yang, Rebecca J. Passonneau, and Gerard de Melo. 2016. PEAK: Pyramid Evaluation via Automated Knowledge Extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2673–2680. AAAI Press.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 563–578. Association for Computational Linguistics.