Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary

Daniel Deutsch,[†] **Tania Bedrax-Weiss,**[‡] **and Dan Roth**[†]
[†]Department of Computer and Information Science, University of Pennsylvania, United States
[‡]Google Research, United States
{ddeutsch,danroth}@seas.upenn.edu
tbedrax@google.com

Abstract

A desirable property of a reference-based evaluation metric that measures the content quality of a summary is that it should estimate how much information that summary has in common with a reference. Traditional text overlap based metrics such as ROUGE fail to achieve this because they are limited to matching tokens, either lexically or via embeddings. In this work, we propose a metric to evaluate the content quality of a summary using question-answering (QA). QA-based methods directly measure a summary's information overlap with a reference, making them fundamentally different than text overlap metrics. We demonstrate the experimental benefits of QA-based metrics through an analysis of our proposed metric, QAEval. QAEval outperforms current state-of-the-art metrics on most evaluations using benchmark datasets, while being competitive on others due to limitations of state-of-the-art models. Through a careful analysis of each component of QAEval, we identify its performance bottlenecks and estimate that its potential upper-bound performance surpasses all other automatic metrics, approaching that of the gold-standard Pyramid Method.[1]

1 Introduction

Evaluating the content quality of a summary is a fundamental task of text summarization. As such, it has received the attention of researchers for the past two decades (Lin, 2004; Nenkova and Passonneau, 2004; Hovy et al., 2006; Louis and Nenkova 2013; Zhao et al., 2019, among others). The most popular approaches are reference-based metrics, which treat a human-written reference summary as the gold standard and score a candidate summary based on how similar its content is to the reference.

It is desirable to have reference-based evaluation metrics that calculate this similarity score based on how much information the two summaries have in common. The vast majority of previous automatic evaluation metrics compare two summaries based on matching their tokens, either through some lexical (Lin, 2004; Hovy et al., 2006; Tratz and Hovy, 2008) or embedding-based similarity (Zhang et al., 2020; Zhao et al., 2019). Although they capture a valuable quality signal, these methods match tokens that do not express the same information and instead end up comparing the similarity of two summaries based on the topics they discuss (Deutsch and Roth, 2020).

In this work, we propose a metric to evaluate the content quality of a summary using question-answering (QA). Metrics within a QA evaluation framework represent the information of a reference summary using QA pairs, then estimate how much of this information is contained in a candidate summary by calculating the proportion of questions it can answer. Because the questions can only be answered if the candidate summary contains the corresponding information, QA-based metrics directly measure the information overlap, providing a summary quality signal that is not effectively captured by text overlap based metrics.

We build upon previous work in this direction (Eyal et al., 2019) and propose and analyze a more general QA-based metric, which we call QAEval (§3.1). We experimentally show the benefit of QAEval, both with current state-of-the-art methods and by estimating its potential upper-bound performance.

[1]Code is available at https://github.com/CogComp/qaeval-experiments.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 774–789, 2021. https://doi.org/10.1162/tacl_a_00397
Action Editor: Xiaojun Wan. Submission batch: 11/2020; Revision batch: 3/2021; Published 8/2021.

We show that with current question-generation and question-answering models, QAEval achieves state-of-the-art correlations to human judgments on benchmark datasets when used to evaluate summarization systems (by averaging scores over dozens of summaries), outperforming all other automatic metrics and equalling the gold-standard Pyramid Method (Nenkova and Passonneau, 2004, §8). When used to rank individual summaries, the metric is equal to or better than other metrics on summaries that are very similar to the ground-truth and is competitive on others due to shortcomings of current state-of-the-art models (§7).

Through a careful analysis of each component of QAEval (§5-§7), we identified 2 performance bottlenecks: (1) the QA model and (2) the task of verifying if the predicted answer is correct (§7), whose noise likely explains the lower summary-level performance in some scenarios. Based on a manually annotated set of 2.9k QA pairs, we show that with human-level QA and answer verification performance, the summary-level upper-bound correlations of QAEval are better than all other automatic metrics and approach the gold-standard Pyramid Method. In combination with state-of-the-art correlation results, this strongly indicates that QA-based evaluation metrics are a promising direction for future research.

The contributions of this work include (1) a proposal of QAEval, a more general QA-based metric for evaluating the content of summaries, (2) experimental evidence that demonstrates QAEval's state-of-the-art performance on benchmark datasets, (3) an analysis that identifies the QA model and answer verification as the performance bottlenecks, and (4) an estimate that QAEval's upper-bound summary-level performance in scenarios in which it currently lags behind is high, approaching that of the gold-standard manual evaluation metric, the Pyramid Method.

2 Related Work

By far the most popular automatic methods for evaluating the content of a summary do so by comparing the tokens of the candidate and the reference. The de facto metric ROUGE (Lin, 2004) calculates a precision and recall score on the summaries' lexical overlap. Recent methods BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019) instead compare tokens based on the similarity of their contextual word embeddings.

Because these text overlap metrics do nothing to specifically measure how much information is common between two summaries, their scores are polluted by spurious matches between tokens that do not express the same information. In contrast, QA-based evaluation metrics *do* directly compare summaries based on their information.

The gold-standard for manually comparing two summaries' information overlap is the Pyramid Method (Nenkova and Passonneau, 2004). It uses a domain-expert to identify spans of text between the candidate and reference summaries that express the same information, known as summary content units (SCUs). Because the Pyramid Method's final score is calculated exclusively on the number of common SCUs, it is a purely information-based evaluation.

While there have been efforts to crowd source the Pyramid Method (Shapira et al., 2019), fully automatic approximations PEAK (Yang et al., 2016) and PyrEval (Gao et al., 2019) have also been proposed, with PyrEval reporting the best performance. PyrEval identifies and matches SCUs by decomposing sentences into clauses, then calculating the similarity of the clauses based on their phrase embeddings. This style of metric has been met with less success than text overlap metrics.

Several recent works also use QA to evaluate summaries. Narayan et al. (2018) use QA as part of a human evaluation to measure how much important document information was maintained by the summary. FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020) automate evaluating the faithfulness of a summary. Faithfulness and content quality are related, yet distinct, concepts. Quality is a measure of whether the summary contains the correct information, whereas faithfulness measures whether the information is consistent with the input, regardless of its importance. FEQA and QAGS compare summaries to the input documents, whereas we compare summaries to references. Because the datasets used in our experiments are extractive summaries or have relatively high faithfulness ratings (Fabbri et al., 2020), we assume faithfulness is not an issue for simplicity.

Then, the most closely related work to ours is Eyal et al. (2019), who also use QA to evaluate the content of summaries via their metric APES.

They create fill-in-the-blank questions by removing named entities from the reference summary and use a reading comprehension model to predict which entity was removed using the candidate summary.

There are several differences between their work and ours. Our proposed metric QAEval is more general than APES because QAEval asks and answers questions about noun phrases, whereas APES is restricted to named entities. APES may fail to accurately score summaries that do not have a sufficient number of named entities. Then, our evaluation of QAEval is more comprehensive: The experiments in Eyal et al. (2019) were limited to evaluating APES on 8 input instances from TAC'11,[2] whereas our experiments are run on 92 instances from benchmark content quality datasets TAC'08 and '09 as well as 100 instances from the CNN/DailyMail dataset (Nallapati et al., 2016; Fabbri et al., 2020). Since our evaluation is more comprehensive and we demonstrate our metric has a high upper-bound performance, we believe it is a more convincing argument that QA-based metrics are a promising direction of future research. Further, we perform an extensive evaluation on the individual components of the metric. We compare our metric's performance to APES' in §8 and §9.

3 QA-Based Evaluation

The standard line of research for evaluating the content quality of a summary is based on comparing the text of a candidate summary to a reference summary. Metrics that follow this approach include ROUGE, Basic Elements (Hovy et al., 2006), AutoSummENG (Giannakopoulos et al., 2008), METEOR (Denkowski and Lavie, 2014), BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), and many more.

It is desirable to evaluate a summary based on the quality of the summary's information. For reference-based metrics, this means measuring the overlap in information between the candidate and reference summary. However, there is evidence that suggests text overlap metrics do not successfully accomplish this (Deutsch and Roth, 2020). They match tokens which do not express the same information and end up comparing the similarity of two summaries based on the topics they discuss.

We argue that a much better method of comparing the information content of two summaries is through QA. In an ideal QA-based evaluation framework, all of the reference summary's information is represented by a set of QA pairs, and the candidate summary's recall of this information is measured by answering the questions against the candidate. The questions should only be answerable if the information necessary to answer them is present in the candidate. Therefore, this approach is fundamentally different from text overlap methods because it explicitly measures how much of the reference's information is contained in the candidate.

While we cannot yet achieve this ideal QA-based metric (our QA-based representations may be incomplete, our QA models are imperfect, etc.), we next propose a specific instantiation of this framework that represents our best effort at reaching this goal with today's state-of-the-art models.

3.1 QAEval

At the core of this work is a reference-based summarization evaluation metric that estimates the content quality of a summary, which we call QAEval. The metric represents the information of a reference summary by a set of question-answer pairs that are automatically generated from the reference. Then, QAEval estimates how much of this information is in a candidate summary by using a learned QA model to answer the questions against the candidate. The predictions from the QA model are verified as correct or incorrect, then the final score of the metric calculates what proportion of the questions were answered correctly.

Below, we describe the individual steps of the evaluation metric in more detail. Then, each component of QAEval is analyzed individually in Sections 5, 6, and 7 in order to identify any performance bottlenecks, followed by an overall evaluation of the metric in Section 8 and a reproduction of the experiments of Eyal et al. (2019) in Section 9.

Answer Selection The first step in generating questions from the reference summary is to pick a set of phrases that represents answers to questions that will later be generated. The answers should be chosen such that they will generate questions that cover as much of the information of the summary as possible. We evaluate how much semantic content is represented by several different answer selection strategies in §5.

[2]https://tac.nist.gov/.

Question Generation Once the answers have been selected, a learned model is used to generate a question for each answer. The input to the question-generation model is a sentence which contains an answer phrase that is demarcated by special tokens. The output is a question which is answerable by that phrase.

Following Durmus et al. (2020), the generation model is a fine-tuned BART model (Lewis et al., 2020) trained on 55k human-written question-answer pairs collected by Demszky et al. (2018). The quality of the generated questions and the impact of using model-generated questions instead of human-written questions on downstream correlations is measured in §6.

Question Answering Given a set of QA pairs generated from the reference summary, a QA model is used to answer the questions against the candidate summary. Since there are no summarization datasets with labeled QA pairs, the QA model must be trained on a different dataset. Further, because it is almost always the case that the candidate summary will not contain some reference summary information, it is necessary for the model to decide whether a question is answerable to reduce noise from spurious answers. Therefore, the QA model is trained on the SQuAD 2.0 dataset (Rajpurkar et al., 2018) which contains unanswerable questions.

The QA model is a pre-trained ELECTRA-Large model (Clark et al., 2020) fine-tuned on SQuAD 2.0. The input to the model is the candidate summary and a question. The output is a span of text that contains the answer or a null string if the question is not answerable, depending on which is more probable under the model. We estimate the answering performance of the QA model on the summarization data and estimate the improvement in downstream correlations that would be expected if the QA model had human-level performance in §7.

Answer Verification and Scoring Finally, once the QA model has output predictions for all of the questions generated from a reference summary, they are verified as being correct or incorrect with respect to the ground-truth answers that were used to generate the questions. We employ the two standard answer verification methods used by SQuAD, exact match (EM) and F_1 (Rajpurkar et al., 2016). If the QA model outputs the null string, the score for that answer is 0. We estimate whether these imperfect answer comparison strategies negatively impact downstream correlations in §7.

Finally, the metric produces two final scores that are the total EM and F_1 scores divided by the number of questions, thus calculating the proportion of questions answered correctly. If multiple reference summaries are available, the scores are macro-averaged. We refer to the metrics as QAEval-EM and QAEval-F_1.

4 Experimental Methodology

We briefly review the experimental methodology that is used to evaluate metrics.

Evaluation metrics are used to estimate some property of a summary that is difficult to directly measure, such as the quality of its content. In order to estimate how well the metric approximates the desired property (i.e., evaluate the evaluation metric), a set of summaries that have been annotated by human judges for that property is scored by the metric, and then the correlation between the two sets of scores is calculated. The summaries are typically the outputs from multiple summarization models for the same set of inputs.

There are two standard ways to calculate correlations in the summarization literature: summary-level and system-level. Assume x_i^j and y_i^j are two scores of metrics X and Y for the summary output by system $i \in \{1, \ldots, N\}$ on input $j \in \{1, \ldots, M\}$. The summary-level correlation is calculated between the scores for each *summary* for the same input, then averaged across inputs:

$$\rho_{\text{SUM}} = \frac{1}{M} \sum_j \textsc{Corr}\left(\left\{ \left(x_i^j, y_i^j \right) \right\}_{i=1}^{N} \right)$$

where $\textsc{Corr}(\cdot)$ calculates some correlation coefficient, typically Pearson r, Spearman ρ, or Kendall τ. Summary-level correlation measures how similarly X and Y rank summaries per-input. In contrast, the system-level correlation is calculated between the scores for each *system* (typically the average score across the inputs):

$$\rho_{\text{SYS}} = \textsc{Corr}\left(\left\{ \left(\frac{1}{M} \sum_j x_i^j, \frac{1}{M} \sum_j y_i^j \right) \right\}_{i=1}^{N} \right)$$

It measures how similarly X and Y rank summarization systems.

In this work, we examine how well evaluation metrics estimate the content quality of a summary using three English summarization datasets: the benchmark TAC'08 and '09 datasets (Dang and Owczarzak, 2008, 2009) as well as the subset of the CNN/DM dataset (Nallapati et al., 2016) that was annotated by Fabbri et al. (2020).

The TAC datasets consist of 48/44 multi-document summarization instances, each with 4 reference summaries written by human annotators. Domain-expert judges rated the summaries output by 58/55 extractive models for each input on a scale of 1 to 5 based on how well they respond to an information need included in the task description. Each summary is also assigned a Pyramid Score (Nenkova and Passonneau, 2004) using a Pyramid constructed from the 4 reference summaries. Our experiments on TAC calculate the correlations of the metrics to the responsiveness score for the 58/55 model summarizers and 48/44 instances.

The annotations provided by Fabbri et al. (2020) on the single-document summarization CNN/DM dataset score the outputs of 16 models across 100 instances. The models are a mixture of extractive and abstractive approaches, and each instance has 1 reference summary. Fabbri et al. (2020) collected relevance scores from 3 expert annotators that captures if the summary contains important content from the input document. Our experiments report the correlation between the metrics' scores and the expert relevance judgments.

5 Answer Selection

In order for a QA-based evaluation metric to be successful, the QA pairs it uses to probe the candidate summary must represent a significant proportion of the reference summary's information. Therefore, in this section, we aim to understand how much information the QA pairs in QAEval do represent and whether that may limit the metric's performance.

We explore three different answer selection strategies which pick phrases that are (1) named-entities, (2) noun phrase chunks, (3) or maximally sized noun phrases. The maximally-sized noun phrases in a sentence are identified by traversing the dependency tree down from the root until a noun is reached, then selecting the entire subtree for that noun. Example answers selected by each strategy are presented in Figure 1.

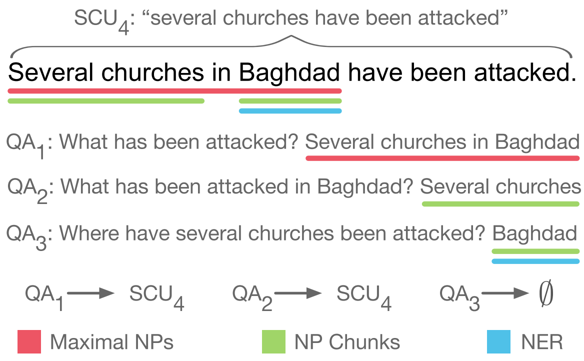

Figure 1: Example answers selected by the three strategies. The *only* SCU marked by annotators for this sentence is SCU_4, which does not include information about the location of the attacks. Therefore, an answer selection strategy that chooses ''Baghdad'' enables generating a QA pair such as QA_3, which probes for information not included in the Pyramid annotation.

Since there is no well-established method of measuring how much semantic content is represented by a set of QA pairs, we instead compare the content covered by the QA pairs to that of another semantic representation, the Pyramid Method SCUs (see §2 for details). This approach allows us to compare answer selection strategies to a common point of reference as well as understand what types of information are represented by each formalism.

In order to compare the content covered by QA pairs and SCUs, each QA pair is manually mapped to an SCU based on whether the information that is being probed by the QA pair is included in the SCU description. For instance, in Figure 1, QA_1 and QA_2 map to SCU_4 because they target what was attacked, which is included in the SCU description, whereas QA_3 would not because the SCU does not describe the location of the attacks. This mapping allows us to calculate the proportion of QA pairs that map to some SCU, called the *QA precision*, and the proportion of SCUs that are mapped to by some QA pair, called the *SCU coverage*.

To ensure that the generated questions are of high-quality, one of the authors manually wrote questions for every answer selected by each strategy for 20 reference summaries across 10 input document sets from TAC'08, totaling 801 questions. The same author further mapped every QA pair to SCUs. The results (averaged over reference summary) are presented in Table 1.

The most significant result we find is that the NP chunks strategy covers 91% of the semantic

Strategy	Avg #QAs	QA Precision	SCU Coverage
NER	11.7	83%	57%
NP Chunks	28.8	79%	91%
Max. NPs	17.3	82%	77%

Table 1: The NP chunks answer selection strategy covers 91% of the information represented by the Pyramid Method (SCU Coverage) with 21% of the questions representing new information. From this, we conclude that the QA pairs generated from selecting noun chunk answers provides a semantic representation of the reference summary with very high-coverage.

information included in the Pyramid Method, with an additional 21% of the questions targeting new information the Pyramid Method does not represent. The other two strategies have much lower SCU coverages, likely because they result in fewer generated questions since their QA precisions are approximately equal to that of NP chunks.

This result is very promising for QA-based evaluation metrics because it indicates that the QA pairs cover nearly all of the information that is used by the Pyramid Method, the best-performing manual content quality evaluation. Further, they even cover information the Pyramid Method does not, suggesting the potential for even better downstream correlations. Therefore, we conclude that the information represented by the QA pairs generated from selecting noun chunk answers is unlikely to be a factor which limits QAEval's performance, and we subsequently use that selection strategy for the rest of our experiments.

Comparing QA Pairs and SCUs Upon comparing the information that is represented by one formalism and not the other, there are some key differences. The QA pairs miss information represented by nominal and adjectival modifiers because that information is contained within the answer noun phrase. For instance, for sentence *[A Turkish novelist] was arrested*, the question asks about who was arrested, and not about the nationality of the novelist, which the SCUs do include.

In contrast, the SCUs often miss specific details and generalize over information that the QA pairs do not. For instance, in Figure 1, although the SCUs do represent that the church attacks happened, it does not include information about their location, whereas this information is targeted by the QAs pairs.

Input: On Jan. 7, 2005, with inauguration scheduled for Jan. 12, [Rossi] filed a lawsuit seeking a new election.

Expert: Who filed a lawsuit seeking a new election?

Model: On Jan. 7, 2005, with inauguration scheduled for Jan. 12, who filed a lawsuit seeking a new election?

Figure 2: A typical example of expert-written and model-generated questions answerable by the phrase in red. The model questions are often significantly more verbose than the expert questions, typically copying the majority of the input sentence.

6 Question Generation

An ideal question generation model should generate questions that are high enough quality that they do not impact the overall performance of the metric. In this section, we compare questions generated by the learned model to expert-written questions, both empirically and extrinsically through downstream correlations to human judgments.

Empirical Analysis Upon comparing the expert-written questions from §5 to model-generated questions for the same set of answers, we observe that a major difference between questions written by an expert versus a model is the level of verbosity. The model-generated questions often copy most of the input sentence over to the question, including parts of the sentence that may not be relevant to answering the question. In contrast, the questions written by an expert are more concise and remove the irrelevant details. Examples of this difference can be seen in Figure 2.

Despite the verbosity, nearly all of the model-generated questions are understandable to the authors. However, because they are rather formulaic, the questions sometimes sound unnatural and could be confusing to a layperson. We did not find any examples in which the answer was included in the question.

Downstream Correlation Ideally, a QA-based evaluation metric would use an expert to write the questions to ensure that they are all high-quality. Unfortunately, this does not scale and is very expensive and time consuming, so the questions must be model-generated. However, it is important to quantify any drop in performance caused by generating questions from a model rather than a domain-expert to understand the impact of using a less-than-ideal approach.

In order to measure any potential drop in performance, we compared the downstream

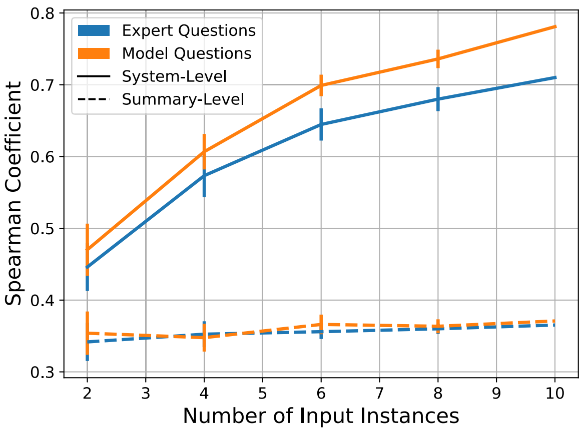

Figure 3: A comparison of the correlations of QAEval-F_1 on a subset of TAC'08 using expert-written and model-generated questions. Each point represents the average correlation calculated using 30 samples of $\{2, 4, 6, 8, 10\}$ instances, plotted with 95% error bars. System-level correlations were calculated against the summarizers' average responsiveness scores across the entire TAC'08 dataset. We hypothesize the model questions perform better due to their verbosity, which causes more keywords to be included in the question that the QA model can match against the summary.

correlations of the QA-based metrics to responsiveness judgments when using expert-written and model-generated questions. In both cases, the question-answering component was done using the learned model described in §3.1.

This experiment was performed on the subset of the TAC'08 dataset for which we collected expert-written questions (see §5). That is, the summaries from 58 different systems across 10 input instances with 2 references each were scored using the two setups, and the respective correlations were computed.[3] We further simulated having a smaller number of input instances by downsampling the data to observe any emerging trends. The results are plotted in Figure 3.

The downstream summary-level correlations appear near-identical between the two approaches. However, surprisingly, the model-generated questions appear to result in better downstream correlations at the system-level than the expert-written questions. As soon as around 6 input instances are available, the two curves separate from each other's margins of error, with the model-generated questions clearly trending with a Spearman correlation of at least 0.05 higher.

It is not clear from examining the data why this is the case. There is no clear pattern that emerges which could explain why the model-generated questions result in higher correlations. Our best hypothesis is that the verbosity of the generated questions helps the QA model by including more keywords that can be matched against the summaries to find an answer.

From these unexpected results, we can conclude that the model-generated questions do not harm the downstream correlations of QAEval at either the summary- or system-levels. The rest of the experimentation in this paper will only use model-generated questions.

7 Question Answering and Verification

The task of the QA model and answer verification step are to determine whether a question is answerable against a summary, predict an answer if it is, then compare the prediction to the ground-truth answer to determine if it is correct. In this section, we evaluate the performance of both components on the summarization data, first by calculating the QA performance (§7.1) and then by estimating the downstream correlation of QAEval if both components had human-level performance (§7.2).

7.1 Question-Answering Model Performance

Since the QA model is trained on Wikipedia articles in the SQuAD 2.0 dataset and used to answer questions generated from the summarization data, it is expected that the QA performance on the summarization data will be worse than on the original training data due to the domain shift.

In order to quantify the size of such a drop, one of the authors manually answered 2.9k generated questions from 20 reference summaries across 10 input clusters against 4 different summarizers on TAC'08 and 2.3k generated questions across 10 input documents against all 16 summarizers on CNN/DM. For each question and summary pair, it was first determined whether the summary contained the answer to the question, then, if it did, a span of text was selected as the answer. Next, the selected answer was later manually verified as correctly or incorrectly answering the question.

We compare the QA model's ability to both identify if a question is answerable and to select the correct answer if one exists separately on SQuAD 2.0 and the summarization datasets. This is done to measure any performance decrease on each

[3] Since we do not have expert-written questions for all 4 references across all 48 input clusters, these results are not strictly comparable to later experiments (e.g., §8).

Summary: The killing of Lebanon's former PM Rafiq Hariri renewed calls for Syria to abide by UN Security Council Resolution 1559 and end its dominance of Lebanon…

Question: What event put Syria under renewed pressure from the international community to abide by UN Security Council Resolution 1559 and withdraw its troops from Lebanon?

Answer: The February assassination

Prediction: The killing of Lebanon's former PM Rafiq Hariri

Figure 4: An example correct answer predicted by the model that is scored poorly by the EM or F_1 QA metrics (both would assign a score of 0 or near 0). This occurs because the answer and prediction are drawn from two different summaries, and the same event is referred to in different ways in each one.

Dataset	%IsAns	IsAns-F_1	Given IsAns		
			EM	F_1	Acc
SQuAD 2.0	50.0%	92.0	88.0	94.5	–
TAC'08	14.2%	52.4	56.5	69.5	84.3
CNN/DM	36.3%	75.3	73.8	83.6	86.3

Table 2: The QA performance on the summarization datasets drops significantly compared to its performance on SQuAD, especially for TAC'08. This is expected due to the domain shift, however we suspect the drop is smaller for CNN/DailyMail because the generated and reference summaries are far more similar than for TAC, thus making it easier to answer questions.

problem in isolation. We calculated the F_1 score on the model's predictions on whether the question is answerable, plus the standard SQuAD EM and F_1 metrics on only the subset of QA pairs for which the ground-truth and model agree that the question is answerable. We do not want to measure the quality of the predicted answer if the question is not answerable or the model outputs no answer.

In addition to EM and F_1, we also report the correct answer accuracy according to the human annotator. EM and F_1 are imperfect answer comparison strategies because they may fail to identify an answer as correct if it is a paraphrase of the ground-truth. Unlike SQuAD, the ground-truth answer and model prediction come from different source texts, increasing the likelihood that both answers will be expressed differently (see Figure 4). Comparing the human annotator accuracy to EM and F_1 will quantify how well the automatic answer verification methods work on the summarization data.

The results are presented in Table 2. In general, the QA performance drops for both datasets, but the decrease is more extreme for TAC'08. Specifically, we see that the drops in IsAns-F_1 are significant, amounting to decreasing by nearly 40 points from 92.0 on SQuAD to 52.4 on TAC'08 and almost 17 points to 75.3 on CNN/DM. This result indicates that identifying if a question is answerable is very challenging for the model, especially on TAC.

The EM and F_1 results across datasets also see a rather significant drop of around 47–58 points for TAC and 42–55 for CNN/DM, pointing to a much worse answering performance by the model when the model correctly predicts that an answer exists. However, the accuracy according to the human annotations is closer to the performance on SQuAD, implying the actual drop in performance is actually not as significant. The discrepancy between the EM/F_1 scores and human accuracy judgments means the model's predictions are frequently correct, but EM and F_1 fail to identify them as such in a significant number of cases, thereby implying they are noisy answer verification methods. This problem has been observed for QA models before (Wang et al., 2020; Chen et al., 2020), but the issue seems particularly apparent for when the answer and QA models prediction come from different texts.

We suspect that the QA model fares better on CNN/DM than TAC because the CNN/DM generated summaries are far more similar to the reference summaries than those in TAC. This is likely due to several factors: (1) The CNN/DM task is in some sense easier than the TAC task. The lead-3 baseline is very strong, so the models can more easily generate high quality summaries. (2) The models included in the annotation are more recent state-of-the-art models compared to those from TAC and are likely better summarizers. (3) The task is single-document, so the information in the reference and generated summaries is more likely to be expressed the same way.

Since the two summaries being compared are similar to each other, the generated questions have a large token overlap with the target summary. This likely results in the QA model being more effective at identifying when an answer exists in the summary and then subsequently correctly identifying it. We expect this result to hold for other popular single-document summarization datasets.

From this experiment, we conclude that identifying whether a question is answerable and subsequently verifying whether the QA models prediction is correct are potential performance bottlenecks QAEval.

7.2 Human-Level Performance Comparison

After identifying QA and answer verification as potentially problematic for QAEval's performance, we now estimate the size of any potential drop in downstream correlation compared to using human-level performance for both of those components.

Using the same human-annotated QA pairs from the previous section, we calculated the summary-level correlations of QAEval when it uses either human annotations for the QA model, human annotations for the answer verification, or both. The correlations for these QAEval variants and several other metrics (discussed in §2) are in Table 3.

Since this experiment only uses a relatively small amount of data, none of the correlations differ by a statistically significant margin, so coming to definitive conclusions is difficult. However, some trends do emerge from the data.

For TAC'08, QAEval is competitive to the other evaluation metrics when it uses a learned QA model and F_1 verification. Then, human-level performance for both QA and answer verification provide large improvements in the downstream correlations, both independently and when combined. For instance, human QA annotations improve QAEval on TAC by 0.12 and 0.14 Pearson with F_1 and human verification, respectively. Human annotations for answer verification improve QAEval with model and human QA components by 0.17 and 0.29 Spearman, respectively. When both components use human annotations, the correlations are significantly better than any of the other automatic metrics and approach those of the Pyramid Method.

The results on CNN/DM are less clear. There is no obvious pattern in the data and all of the model/human combinations result in roughly the same performances. We suspect that because the drop in QA performance is less significant (§7.1), the differences in model and human-level QA performance is not reflected on CNN/DM as it is on TAC. Further, we empirically observed that the content of this dataset's summaries are more similar in content across models than the TAC summaries, making them harder to rank (as demonstrated by the lower correlations), which would also introduce more variance to the correlations.

System		**TAC'08**			**CNN/DM**		
		r	ρ	τ	r	ρ	τ
Pyramid Score		.63	.69	.65	–	–	–
ROUGE-1		.27	.27	.26	.25	.21	.18
ROUGE-2		.34	.40	.38	.13	.09	.06
ROUGE-L		.20	.22	.21	.13	.12	.08
ROUGE-SU4		.29	.22	.22	.16	.16	.12
MoverScore		.42	.28	.28	.27	.23	.18
APES		.35	.38	.37	.08	.09	.07
QAEval							
QA	Verif.						
Model	F_1	.31	.28	.26	.21	.23	.18
Human	F_1	.43	.33	.30	.15	.14	.12
Model	Human	.44	.45	.42	.25	.24	.20
Human	Human	.58	.62	.59	.22	.21	.17

Table 3: Summary-level correlations calculated using 4 systems across 10 inputs on TAC and 16 systems across 10 inputs on CNN/DailyMail compared using answers from a model or a human and verifying if the answer is correct using F_1 or a human. Because the results are on a small sample of the dataset, the results are not statistically significant. However, the trend on TAC is that human-level performance greatly improves the results, approaching correlations equal to the Pyramid Method's. On CNN/DailyMail, we suspect the same trend does not appear because the QA model performs much better than on TAC.

Overall, this is a promising result for the future potential of QA-based evaluations, especially for more complex multi-document summarization tasks which are in some sense harder for metrics to evaluate than single-document summaries. While the current summary-level results on both datasets may be competitive to other metrics, the metric's upper-bound performance is very high on TAC and is approaching the gold-standard manual evaluation, the Pyramid Method.

8 Overall Metric Analysis

After analyzing each component of QAEval, we now turn to calculate the metric's correlations

TAC 2008							TAC 2009						
Metric	System-Level			Summary-Level			Metric	System-Level			Summary-Level		
	r	ρ	τ	r	ρ	τ		r	ρ	τ	r	ρ	τ
Pyramid Score	.90	.88	.70	.59	.59	.50	Pyramid Score	.90	.87	.70	.59	.57	.48
ROUGE-1	.79	.80	.60	.49	.48	.39	ROUGE-1	.83	.78	.60	**.54**	.47	.38
ROUGE-2	.83	.87	.67	.48	.48	.39	ROUGE-2	.76	.84	.67	.50	.50	.40
ROUGE-L	.74	.77	.57	.46	.45	.36	ROUGE-L	.82	.72	.54	**.54**	.47	.37
ROUGE-SU4	.80	.83	.63	.49	.48	.39	ROUGE-SU4	.77	.81	.63	.52	.50	.39
PyrEval	.81	.79	.59	.31	.31	.25	PyrEval	.86	.82	.64	.39	.35	.28
MoverScore	.83	.82	.63	**.50**	**.49**	**.40**	MoverScore	.82	.80	.63	.51	**.52**	**.42**
APES	.74	.82	.60	.25	.25	.21	APES	**.87**	.80	.63	.41	.35	.28
QAEval-EM	**.93**	**.91**	**.76**	.33	.33	.27	QAEval-EM	.70	.87	.69	.42	.38	.30
QAEval-F_1	.90	.88	.71	.46	.45	.36	QAEval-F_1	.81	**.89**	**.72**	.50	.45	.36

Table 4: The Pearson r, Spearman ρ, and Kendall τ correlation coefficients calculated between the metrics' scores and expert responsiveness judgments on the TAC'08 (left) and TAC'09 (right) datasets. QAEval has the highest system-level correlations, even better than the fully manual Pyramid Score, whereas the summary-level correlations are lower (EM) or competitive (F_1) with other metrics. We believe this supports our hypothesis that the QA model and answer verification are noisy (causing lower summary-level correlations) but average out to a high-quality metric given enough QA pairs (causing high system-level correlations). On TAC'09, the QA r values are much lower because of an outlier, and r is sensitive to outliers. If the outlier is removed, the r values become 0.92 and 0.93 for EM and F_1.

to human responsiveness/relevance judgments on TAC'08, '09, and CNN/DM (see §4 for more details about the experimental methodology; an additional experiment that varies the number of available references is included in Appendix A). For this experiment, QAEval uses the NP chunks answer selection strategy and learned question-generation and question-answering models and is therefore a fully automatic metric.

In addition to the QAEval correlations, we also report those of several baselines and state-of-the-art metrics, including the Pyramid Score, several variants of ROUGE, PyrEval, and MoverScore (which reports better correlations than BERTScore), and APES. See §2 for descriptions of these metrics. Results in bold are the highest among the automatic metrics. Those underlined are statistically indistinguishable from the highest under a single-tailed permutation test for correlations with $\alpha = 0.05$ (Deutsch et al., 2021).

TAC'08 and '09 The correlations for TAC are presented in Table 4. First, we see that the summary-level correlations for the QAEval metrics are lower than or comparable to some of the other automatic metrics. For example, the TAC'08 Pearson's r for QAEval-EM is 0.33, whereas the r values for QAEval-F_1 and ROUGE-2 are 0.46 and 0.48. Given that the QA model and answer verification components introduce noise into the metric, this result is consistent with the analysis in §7.2 and unsurprising.

However, the system-level results are quite surprising. The QAEval metrics achieve state-of-the-art system-level performance on nearly every correlation coefficient across both datasets, reaching correlations comparable to the Pyramid Method itself. For instance, on TAC'08, QAEval-EM has a Kendall's τ of 0.76 compared to 0.70 for the Pyramid Method and 0.67 for the next-highest automatic metric, ROUGE-2. This pattern largely holds for TAC'09, with the exception of Pearson's r due to an outlier.[4]

It is unexpected that QAEval should achieve both state-of-the-art system-level results and lower summary-level results simultaneously and that the system-level results are even better than the Pyramid Method's.

We believe the discrepancy between the summary- and system-level results can be explained by the number of questions that is used by each evaluation. QAEval estimates the quality of an individual summary using around 110 questions. In contrast, the system-level scores are based over 5,000 QA pairs across 48 or 44

[4]Once removed, the r values are 0.92 and 0.93 for QAEval-EM and QAEval-F_1, higher than any other metric.

Fabbri et al. (2020)						
Metric	System-Level			Summary-Level		
	r	ρ	τ	r	ρ	τ
ROUGE-1	.61	.62	.50	.28	.26	.20
ROUGE-2	.64	.60	.43	.23	.19	.14
ROUGE-L	.61	.48	.32	.21	.18	.14
ROUGE-SU4	.62	.56	.38	.23	.19	.15
MoverScore	.56	.54	.42	.28	.24	.18
APES	.68	.73	.58	.10	.09	.07
QAEval-EM	.80	**.91**	**.77**	.23	.23	.19
QAEval-F_1	**.82**	**.91**	**.77**	**.30**	**.29**	**.22**

Table 5: The QAEval metrics on the CNN/ DailyMail annotations provided by Fabbri et al. (2020) achieve significantly higher correlations than the other authomatic metrics, likely due to the relatively good QA model performance on this dataset compared to on TAC.

instances. We suspect that when QAEval's scores are averaged over such a large number of questions, the metric is able to overcome any noise introduced by the QA model or answer verification, resulting in a high-quality evaluation. APES, the other QA-based metric, also exhibits a similar pattern, supporting this hypothesis.

Then, it is likely that QAEval's system-level performance rivals the Pyramid Method's because the QA pairs probe for more semantic content than is represented by the SCUs (§5). The QA model and answer verification largely perform the same task as the Pyramid Method annotators: identify a span of text in the candidate summary which expresses a specific piece of information. It is that unlikely the models do this better than a human, even after the noise is averaged out across thousands of examples. Therefore, it must be the case that the semantic representation of the QA pairs provides better coverage of the reference summary than the SCUs do, resulting in comparable overall performance.

CNN/DM The results on the CNN/DM dataset are shown in Table 5. Compared to TAC, the improvement in system-level correlations is significantly larger. For instance, both QAEval variants achieve a system-level Spearman 0.91, whereas the next highest metrics APES and ROUGE-1 reach 0.73 and 0.62. Unlike for TAC, the summary-level correlations are either higher or statistically indistinguishable from the other metrics.

We hypothesize that the improved performance on CNN/DM compared to TAC is due to the QA model's quality on this dataset. In §7.1, we demonstrated that the QA performance did drop on CNN/DM with respect to the model's results on the SQuAD data, however that performance decrease was not nearly as large on CNN/DM as on TAC. Since the QA model and answer verification are the performance bottlenecks and both suffer less on CNN/DM, the QAEval metrics achieve strong correlations.

This result is evidence to support that QAEval is a very effective metric for evaluating current state-of-the-art systems on today's popular summarization datasets.

Comparison to APES Across all three datasets, QAEval achieves higher or comparable correlations than the other QA-based metric, APES, at both the summary- and system-levels. We suspect this is due to at least two reasons. First, their reading comprehension model likely has lower performance than the QA model used in QAEval. The QAEval pretrained model leverages recent state-of-the-art models that use contextual word embeddings, which the model of Eyal et al. (2019) does not use. Second, APES targets named entities in the summaries, which we demonstrated does not probe for as much information as using all noun phrases (§5). If the summaries do not contain a sufficient number of entities, APES may fail to accurately score it.

Overall Since the performance of QAEval using EM and F_1 is roughly equal at the system-level, but F_1 is clearly better at the summary-level, we recommend that future work that evaluates with QAEval use the F_1 variant.

Overall, since evaluation metrics are most commonly used in the summarization community to rank summarization systems, these experimental results suggest that QAEval is one of the most effective evaluation metrics to date.

9 APES Experiments

To further compare QAEval to APES, we reproduce some of the experiments reported by Eyal et al. (2019) and compare the results of the two metrics.

9.1 TAC 2011 Comparison

First, we compare the summary-level correlations of the two metrics and ROUGE to human

	R1	R2	RL	RSU4	APES	QA-EM	QA-F1
Pyr.	.73	.73	.70	.74	.61	.47	.61
Resp.	.62	.65	.60	.63	.50	.46	.56

Table 6: Summary-level Pearson correlations of ROUGE, APES, and QAEval to overall responsiveness and the Pyramid Score on the 8 instances from TAC'11 that were used in Eyal et al. (2019). These numbers differ from those reported by Eyal et al. (2019) because they directly calculate the correlation between the scores for all of the summaries across all instances (personal communication with the authors). This differs from the standard definition of the summary-level correlation, which calculates a correlation per input document set (Louis and Nenkova, 2013), then averages the correlations (Peyrard et al., 2017; Zhao et al., 2019; Bhandari et al., 2020, see ρ_{SUM} in §4).

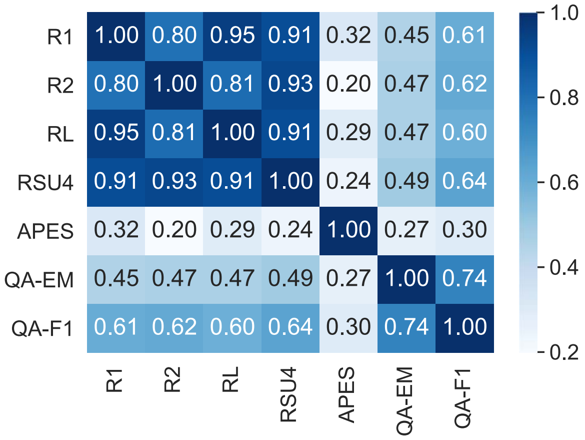

Figure 5: The Pearson correlations between the scores of several ROUGE variants, APES, and QAEval variants on TAC'08. The results support similar findings of Eyal et al. (2019), namely, that the ROUGE metrics are highly correlated to each other but have low correlation to the QA-based metrics, suggesting the two types of metrics offer complementary signals.

judgments on a subset of the TAC'11 dataset. TAC'11 contains extractive summaries produced by 51 models on 44 input document sets. However, Eyal et al. (2019) evaluate on the 8 input document sets about ''Investigations and Trials'' for which there were a sufficient number of named entities. This is because the QA model used by APES is only trained to predict named entities as answers. Similarly to TAC'08 and '09, each summary has an overall responsiveness score and a Pyramid score that were annotated by domain experts.

Table 6 contains the summary-level Pearson correlations of ROUGE, APES, and QAEval to the human judgments on the subset of TAC'11. Although it is difficult to come to conclusions on this dataset due to its relatively small size, we observe that APES out-performs QAEval-EM and under-performs QAEval-F_1 using the responsiveness score as the ground-truth. Using the Pyramid score as the ground-truth, APES and QAEval-F_1 are equal. However, both QA-based metrics are lower than the ROUGE variants, which is consistent with both APES and QAEval achieving lower correlations than ROUGE on TAC'08 and '09 at the summary-level. The APES correlations here are much higher on this subset of TAC'11 than on the whole of TAC'08 and '09, supporting that its performance is higher when the summaries have a sufficient number of named entities.

9.2 Complementary Signals

Then, Eyal et al. (2019) demonstrate that APES and ROUGE are less correlated to each other than ROUGE variants are to themselves, suggesting they offer complementary signals of summary quality. In Table 5 we show the Pearson correlations between several different variants of ROUGE, APES, and QAEval on the TAC'08 summaries.

Our results suggest similar conclusions to Eyal et al. (2019). Specifically, each of the ROUGE variants is very highly correlated to each other ($\geq .80$), whereas the correlations to the QA-based metrics are lower ($\approx .47$ for QAEval-EM, .62 for QAEval-F_1, and .26 for APES). Interestingly, APES and QAEval are as correlated to each other as APES is to ROUGE. We hypothesize that because the QA models are trained on different corpora (CNN for APES versus Wikipedia for QAEval), they learn different signals to answer questions and are more effective at scoring different summaries. Future work could explore combining lexical overlap and QA-based methods into a single metric.

10 Discussion

Limitations Overall, QAEval is limited by its dependence on using predicate-argument relations throughout each component of the metric.

QAEval represents summaries with QA pairs that target nouns as answers, which is insufficient for representing all of the summary's information (as pointed out in §5). The question generation model is limited to producing questions that reason about the arguments of predicates and cannot generate more abstract questions (e.g., *What types of conflict have there been?* for Figure 1). Likewise, QA models trained on SQuAD-style questions can only reason about matches between predicate-argument relations and cannot answer more abstract questions even if the generation model could produce them.

Because of this dependence on predicate-argument relations, any similarity between summaries that cannot be represented by matching predicates and arguments can also not be captured by QAEval. Although this does not appear to be an issue in our experiments, we anticipate that using generation and answering models which are capable of a more sophisticated level of reasoning will be necessary in the future.

QA-Based versus Text Overlap Although QAEval has superior or comparable system-level correlations on the datasets included in our experimentation, it still lags behind text overlap-based method ROUGE at the summary-level in some settings. Therefore, we do not recommend completely replacing text overlap metrics with QAEval, nor do we believe that this should be done even if a QA-based metric achieves summary-level parity.

Both Eyal et al. (2019) and our work clearly show evidence that QA-based metrics provide a summary quality signal that is complementary to ROUGE (§9.2), yet both ROUGE and QAEval achieve strong correlations in our experiments. The quality signals captured by these metrics are clearly both valuable and different. Evaluating a summarization model with only one type of metric would miss out on summary quality signals captured by the other. Therefore, we recommend future work use both a text overlap metric as well as a QA-based metric to evaluate their summarization models.

11 Conclusion

In this work, we proposed a QA-based evaluation metric called QAEval. We demonstrated that QAEval already achieves state-of-the-art system-level correlations, and we estimate its upper-bound summary-level performance on multi-document summaries is quite high. Through a careful analysis of each component of QAEval, we identified that the performance bottlenecks are both the QA model and verifying whether or not the QA model's predicted answer is correct. We believe that these results are strong evidence that QA-based evaluation metrics are a promising direction for future research on summarization evaluation.

Acknowledgments

The authors would like to thank Annie Louis, Shashi Narayan, Yang Gao, Fei Liu, and the anonymous TACL reviewers for their very detailed and helpful feedback on our paper, which we feel significantly strengthened our work.

This work was supported 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.

This research is supported by a Focused Award from Google.

References

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.emnlp-main.751`

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*, pages 6521–6532. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.emnlp-main.528`

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA:

Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net.

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*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *CoRR*, abs/1809.02922v2.

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. https://doi.org/10.3115/v1/W14-3348

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A Statistical analysis of summarization evaluation metrics using resampling methods. *CoRR*, abs/2104.00054v1.

Daniel Deutsch and Dan Roth. 2020. Understanding the extent to which summarization evaluation metrics measure the information quality of summaries. *CoRR*, abs/2010.12495v1.

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, ACL 2020, Online, July 5–10, 2020*, pages 5055–5070. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.454

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. 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 3938–3948. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2020. SummEval: Re-evaluating summarization evaluation. *CoRR*, abs/2007.12626v3.

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, Vangelis Karkaletsis, George A. Vouros, and Panagiotis Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing*, 5(3):5:1–5:39. https://doi.org/10.1145/1410358.1410359

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).

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, ACL 2020, Online, July 5–10, 2020*, pages 7871–7880. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

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.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300. https://doi.org/10.1162/COLI_a_00123

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. https://doi.org/10.18653/v1/K16-1028

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. 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 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1158

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.

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. https://doi.org/10.18653/v1/W17-4510

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, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2124

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.

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. https://doi.org/10.18653/v1/2020.acl-main.450

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. https://doi.org/10.18653/v1/D19-1053

A Number of Available References

Previous work has argued that multiple reference summaries are necessary for metrics to achieve stable correlations to ground-truth annotations, especially at the summary-level (Nenkova and Passonneau, 2004; Louis and Nenkova, 2013). Since the TAC datasets provide four reference summaries per input, we are able to measure how much benefit the additional references provide by simulating having fewer references.

In order to simulate only having one reference summary, for each input document set from TAC'08, we randomly sample one reference, score all of the peer summaries against only that reference, and calculate the correlation to the responsiveness scores. We collect 30 such samples and report the average correlation. This procedure is also repeated for two and three references to understand the impact of each additional reference. The results are plotted in Figure 6.

At the system level, the Pearson correlations are largely the same when the metrics are provided with one or four references. This is in agreement with Louis and Nenkova (2013), who show system-level results are relatively stable with either one or four references. Among the metrics, the QA-based metrics see the largest improvement in performance with adding additional references. QAEval-F_1 increases from 0.85 with one reference compared to 0.90 with four. Despite its drop in performance with one reference, QAEval-F_1 is still better than ROUGE even with four references at 0.79. APES improves from 0.66 to 0.74.

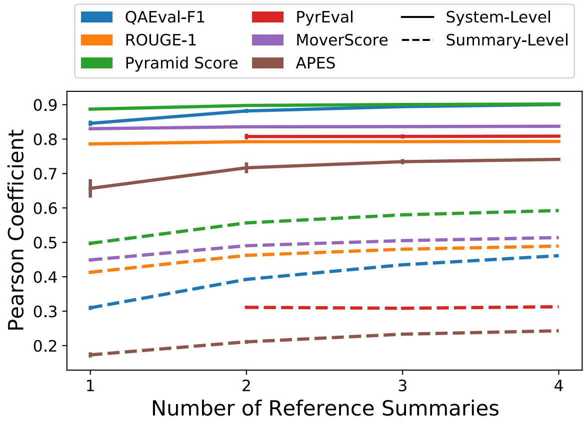

Figure 6: The system- and summary-level Pearson correlations as the number of available reference summaries increases. 95% confidence error bars shown, but may be too small to see. PyrEval is missing data because the official implementation requires at least two references. Even with one reference summary, QAEval-F_1 maintains a higher system-level correlation than ROUGE.

When the metrics are compared at the summary level, it is clear that the correlations for each metric are less stable. Nearly all of the metrics greatly benefit from additional references: Pyramid Score improves by 0.09 (+19%), ROUGE by 0.08 (+18%), and QAEval-F_1 by 0.15 (+49%). The large improvement by QAEval-F_1 is further evidence that the noisy question-answering model averages out to a high-quality responsiveness estimator when provided with a large number of QA pairs.

Overall, QAEval does incur a significant performance drop at the summary-level, but since most comparisons of summarization systems are done at the system-level, it does not appear that having multiple references per input is necessary for good results.