Benchmarking Answer Verification Methods for Question Answering-Based Summarization Evaluation Metrics

Anonymous ACL submission

Abstract

Question answering-based summarization evaluation metrics must automatically determine whether the QA model's prediction is correct or not, a task known as answer verification. In this work, we benchmark the lexical answer verification methods which have been used by current QA-based metrics as well as two more sophisticated text comparison methods, BERTScore and LERC. We find that LERC out-performs the other methods in some settings while remaining statistically indistinguishable from lexical overlap in others. However, our experiments reveal that improved verification performance does not necessarily translate to overall QA-based metric quality: In some scenarios, using a worse verification method – or using none at all – has comparable performance to using the best verification method, a result that we attribute to properties of the datasets.

1 Introduction

A recent trend in summarization metrics is evaluating the quality of a summary via question answering (QA; Eyal et al., 2019; Scialom et al., 2019, 2021; Durmus et al., 2020; Wang et al., 2020; Deutsch et al., 2021a). These metrics compare the semantic content of two texts (e.g., the reference and candidate summaries) by generating questions from one and answering those questions against the other. The amount of common semantic content is proportional to the number of questions which are answered correctly.

A critical step of QA-based evaluation metrics is to verify whether the QA model's prediction is correct, a task known as answer verification (see Fig. 1). This helps to both suppress noisy output from the QA model as well as identify inconsistent information across the texts.

Answer verification is typically done by comparing the prediction to the expected answer by the exact match or token F_1 string comparison methods (Rajpurkar et al., 2016). However, more sophisticated text comparison methods have been proposed in recent years, and it is unknown whether they provide a benefit in this particular scenario.

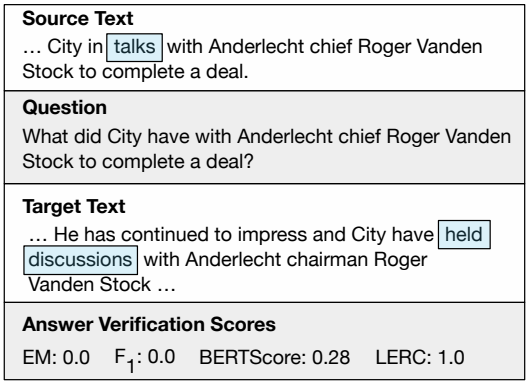

Figure 1: In the answer verification task, the evaluation metrics score how likely two phrases (one the ground-truth answer and one the QA model's prediction) from different contexts have the same meaning.

In this work, we benchmark various answer verification strategies for QA-based summarization evaluation metrics. Our goal is to understand whether methods that are more advanced than lexical overlap are better able to classify phrases as having the same or different meaning as well as whether any such improvements result in the overall QA-based metric being better at replicating human judgments of summary quality.

We analyze four answer verification methods, exact match, token F_1, BERTScore (Zhang et al., 2020), and LERC, (Chen et al., 2020) in combination with two QA-based metrics, QAEval (Deutsch et al., 2021a) and FEQA (Durmus et al., 2020).

Based on a set of human annotations across two datasets, we find that LERC, in general, performs the best at the actual task of answer verification, although in some settings it is statistically indistinguishable from token F_1 (§4.1). However, our results also show that any such improvement in verification performance does not always translate

to a better QA-based evaluation metric (§4.2).

We believe these results can be explained by properties of the QA metrics and the datasets. When the QA model performance is high or the verification task is in some sense easy to do, it may not be necessary to have a sophisticated verification method or even use one at all. Despite this, our recommendation is still to do answer verification with LERC as it can only improve performance, although token F_1 may suffice in some situations.

2 Related Work & Background

The majority of summarization evaluation metrics can be viewed as estimating how similar in meaning two pieces of text are. For instance, ROUGE (Lin, 2004) does this by calculating the number of overlapping n-grams between the two texts.

Instead of directly comparing the entire texts, QA-based metrics identify specific phrases within the texts which should be compared, as follows. First, a set of questions is automatically generated from one text. Then, those questions are automatically answered against a second text to obtain a set of predicted answers. The final score is proportional to the number of correct predictions, but determining whether those predictions are correct – the task of answer verification – is done by comparing the text of the prediction to the expected answer. Therefore, instead of directly comparing the entire contents of the two texts, QA-based metrics instead reduce the scope of the problem to only comparing specific pairs of phrases.

Current QA-based metrics perform the answer verification step by lexical comparison, either exact match or token F_1. Such metrics include QAEval (Deutsch et al., 2021a), FEQA (Durmus et al., 2020), and more (Eyal et al., 2019; Wang et al., 2020; Scialom et al., 2019, 2021). However, any such function which calculates the similarity of arbitrary text can be used instead. This includes embedding-based methods such as BERTScore (Zhang et al., 2020) or metrics which have been trained specifically to do this task, such as LERC (Chen et al., 2020). Evaluating how these methods perform as answer verification methods for QA-based metrics compared to the lexical baselines is the scope of this work.

3 Definitions & Methods

We define the answer verification task as the following: Given a question, answer, the source text from which the QA pair was generated, a prediction, and the target text the prediction comes from, score how similar the meanings of the answer and prediction are (see Fig. 1 for an example).[1] Answer verification is used by QA-based metrics to suppress noisy outputs from the QA model as well as identify when the QA prediction is correct with respect to the target text but incorrect with respect to the expected answer (e.g., unfaithful information).

We analyze four different answer verification methods.

Exact Match The exact match (EM) method compares the two phrases to see if they are identical (after light normalization). EM assigns a score of 1 if the phrases are identical and 0 otherwise.

Token F_1 The token F_1 comparison calculates an F_1 score based on the number of unigrams the two phrases have in common. This is equivalent to the F_1 variant of ROUGE-1.

BERTScore BERTScore (Zhang et al., 2020) compares two pieces of text by aligning the texts' tokens according to which pairs have the highest BERT embedding cosine similarity. We adapt BERTScore to answer verification by encoding the answer and prediction using their respective contexts, then calculating the BERTScore only between the two phrase encodings. Since the output of BERTScore is often in a narrow range of values, we rescale the scores by defining 0 and 1 as the 2.5th and 97.5th percentiles of the BERTScores calculated over the whole dataset.

LERC Chen et al. (2020) proposed LERC, a learned metric for scoring how similar the expected and predicted answers to a question are conditioned on the question and the target text the prediction comes from. All of the inputs are jointly encoded using a BERT-based classifier, which was fine-tuned on human annotations of meaning similarity. Because it was designed for scoring reading comprehension predictions, it does not use the source text. We rescale the output from LERC to be in the range $[0, 1]$.

[1] This is slightly different from the task defined by Chen et al. (2020) which does not include the source text because no such text exists in the standard definition of the reading comprehension task. However, we include it because the source text can be used to create a representation for the answer which may be better than using the question alone.

4 Experiments

The answer verification methods are evaluated independently (§4.1) as well as in combination with two QA-based metrics (§4.2), QAEval (Deutsch et al., 2021a) and FEQA (Durmus et al., 2020). QAEval measures the content quality of a summary (does the summary contain "summary-worthy" information) by using a reference summary as the source text and candidate summary as the target text. In contrast, FEQA estimates the faithfulness of the summary (does the summary contain information consistent with the input) by using the candidate summary as the source text and the input document as the target text.

The experiments are run on two datasets, TAC'08 (Dang and Owczarzak, 2008) and SummEval (Fabbri et al., 2021). These datasets have summaries generated by 58 and 16 models for 48 and 100 inputs, respectively, which are annotated with expert judgments. Both QAEval and FEQA are evaluated on SummEval because it contains annotations for both summary quality and faithfulness, whereas only QAEval is evaluated on TAC'08 since it does not have faithfulness judgments.[2]

4.1 Answer Verification Performance

First, we examine how well each answer verification method accurately scores manually labeled answer pairs from the summarization datasets. For each QA metric and dataset combination, we ran the metric on the summaries, then randomly sampled 200 QA predictions (making 600 total). Each prediction and expected answer were manually annotated by the authors for whether or not the two phrases share the same meaning.

Ideally, the answer verification methods should both successfully classify phrases based on their meaning as well as provide a score close to 1 for phrases with the same meaning and close to 0 with different meanings. These properties are quantified by the binary classification accuracy (assigning labels based on a threshold which maximizes this score) as well as the mean squared error (MSE) of the predicted scores, show in Table 1.

We find that LERC is the only method with the best (or tied for the best) performance across all three metric-dataset combinations. Despite LERC's significant improvement on the SummEval data with QAEval predictions, it is statistically indistinguishable from F_1 on the same dataset with FEQA predictions. We believe this can be explained by which texts are being compared for each metric. FEQA compares the generated summary to the input document. Recent summarization models are known to copy heavily from the input with little high-level abstraction or rephrasing, so comparing phrases with token F_1 is likely to be quite successful. In contrast, QAEval compares the reference and generated summaries. The reference summaries are written by humans, and thus more likely to contain information from the input document which is expressed differently. In such a scenario, the learned metric, LERC, shows strong improvements over F_1.

In general, we find that when BERTScore and LERC do improve over F_1, they do so by identifying paraphrases that have no tokens in common, which sometimes requires world knowledge. Examples of this are included in Appendix B.

Ans. Verif.	QAEval				FEQA	
	TAC'08		SummEval		SummEval	
	Acc	MSE	Acc	MSE	Acc	MSE
Majority Cls	51.5	.49	78.5	.22	56.5	.44
EM	64.5	.36	78.5	.46	76.0	.24
F_1	84.0	.19	79.5	.25	**91.0**	.10
BERTScore	81.0	.16	79.5	.20	82.5	.16
LERC	**85.0**	**.13**	**88.0**	**.11**	88.5	**.09**

Table 1: The binary accuracies and mean squared errors of the answer verification methods evaluated on three metric-dataset combinations with 200 manually labeled examples each. Underlined values are statistically indistinguishable from those in bold under a single-tailed pairwise permutation test with $\alpha = 0.05$.

4.2 Overall Metric Evaluation

Next, we investigate whether the differences in classification performance of the verification methods translate to downstream improvements in the overall quality of the QA-based metrics. To do so, we evaluate different variants of the metrics that use each answer verification method. For both QAEval and FEQA, the final score for the summary is the output of the answer verification method averaged over all of the QA pairs.[3]

QAEval For QAEval, we report the standard system- and summary-level correlations of the metrics' scores to human judgments in Table 2 (due to space constraints, we refer the reader to Deutsch

[2] Our code will be released after publication.

[3] QAEval can also predict a question is unanswerable. In such cases, the score of the prediction is 0.

et al. (2021b) for definitions of the correlations). We also compare against the standard BERTScore and ROUGE metrics as well as a QAEval variant which uses no answer verification by always marking the phrases as correct if the QA model predicts the question is answerable, denoted QAEval-IsAns.

In general, all of the answer verification methods work comparably well, although BERTScore and LERC do statistically improve over the lexical methods in some settings, but not by large margins. We believe the performance of QAEval-IsAns offers an explanation as follows.

Answer verification is not necessary if the QA model is perfect and the summaries are faithful (i.e., the QA prediction is always correct). For SummEval, Deutsch et al. (2021a) demonstrated that QAEval's QA performance was reasonable, and the summaries are very faithful with an average consistency score of 4.7 / 5 according to Fabbri et al. (2021). Therefore, it may be difficult to demonstrate an improvement with any answer verification method even if it is high quality since the need for answer verification is low. Indeed, we see QAEval-IsAns statistically ties the best methods.

On TAC'08, we expect it should be easier to show answer verification helps since Deutsch et al. (2021a) showed the QA performance is poor, suggesting answer verification could help to suppress noisy predictions. Indeed, we do see QAEval-IsAns is statistically out-performed by the verification methods. We suspect the improvements are larger at the system-level than the summary-level because the system quality is estimated over a larger number of QA pairs than an individual summary's quality is. A larger number of questions reduces any noise introduced by the verification methods, resulting in a more accurate estimate of summary quality and a better metric.

FEQA We report the direct correlations between the FEQA metrics' scores and human judgments across all of the summaries in Table 3, including those for ROUGE, BERTScore, as well as FactCC (Kryscinski et al., 2020). FactCC is a learned model to predict the factual consistency between two texts that was trained on synthetically generated data.

Among the FEQA variants, F_1 is the best or indistinguishable from LERC. This result is expected given how similarly they perform at answer verification on this QA metric and dataset split. This is again likely due to the fact that the summarization models copy heavily from the input documents, so the expected answers and QA model predictions are likely to be quite lexically similar. Overall, the FEQA correlations are still lower than those by FactCC by a large margin.

Metric	TAC'08		SummEval	
	Sys	Sum	Sys	Sum
BERTScore	.68†	.40†	.75†	.27†
ROUGE-1	.60	.39†	.50	.20
ROUGE-2	.67	.39†	.43	.14
QAEval-IsAns	.63	.37	.70†	**.26**†
QAEval-EM	**.74**†	.29	.77†	.19
QAEval-F1	.68	.36	.77†	.22
QAEval-BERTScore	.68†	.38†	.77†	**.26**†
QAEval-LERC	.68†	**.39**†	**.80**†	.24†

Table 2: System- and summary-level Kendall's τ (results with Pearson and Spearman are included in Appendix A). Underlined QAEval values are statistically indistinguishable from the best in bold. Values marked with † are statistically indistinguishable from the best metric overall. Statistical testing done using the single-tailed PERM-BOTH permutation test (Deutsch et al., 2021b) with $\alpha = 0.05$.

Metric	r	ρ	τ
ROUGE-1	.13	.13	.11
ROUGE-2	.25	.25	.19
BERTScore	.17	.17	.14
FactCC	.34†	.36†	.29†
FactCCX	.29	.31	.24
FEQA-EM	.17	.14	.11
FEQA-F_1	**.20**	**.16**	**.13**
FEQA-BERTScore	.15	.12	.10
FEQA-LERC	.18	.15	.12

Table 3: The Pearson r, Spearman ρ, and Kendall τ correlations on the SummEval dataset. Values in bold are the best FEQA variants with those underlined being statistically indistinguishable. † marks the best results across all metrics.

5 Conclusion

In this work, we benchmarked four different answer verification methods for QA-based summarization evaluation metrics. Although we were able to identify some methods perform better than others at verification, any such improvement does not necessarily translate a better overall metric quality. We hypothesize that several factors, including the quality of the QA model and properties of the datasets, likely explain this result. Despite this, our recommendation is that practitioners use LERC, although token F_1 may be sufficient in some scenarios.

References

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)*, pages 6521–6532, Online. Association for Computational Linguistics.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. In *Proc. of the Text Analysis Conference (TAC)*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021a. 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.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021b. A Statistical Analysis of Summarization Evaluation Metrics using Resampling Methods. *Transactions of the Association for Computational Linguistics*, 9.

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.

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.

A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, R. Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

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.

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.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization Asks for Fact-Based Evaluation. *arXiv preprint arXiv:2103.12693*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

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.

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.

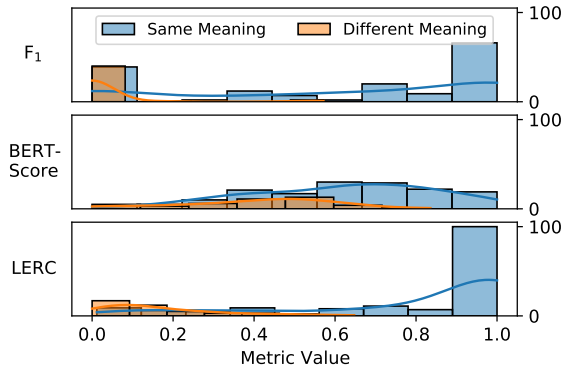

Figure 2: The distributions of score values for three metrics on the SummEval dataset for ground-truth answer and QA model prediction pairs from QAEval with the same (blue) and different (orange) meanings.

A Additional Results

Fig. 2 contains the distributions of score values for token F_1, BERTScore, and LERC on the SummEval dataset grouped by phrases that have and do no have the same meaning. LERC most confidently separates the positive and negative examples. F_1 performs similarly, except it fails in a large number of cases when the two phrases have no tokens in common. BERTScore tends to mix the scores of the positive and negative classes, although they are separated on average.

In Table 4, we report the system- and summary-level correlations on TAC'08 and SummEval with Pearson's r and Spearman's ρ correlation coefficients in addition to the Kendall's τ which was presented in the main body of the paper. The other coefficients lead to a similar conclusion to that which we made with Kendall's τ: All answer verification methods perform comparably well, and when BERTScore or LERC does improve over a lexical baseline, it is not by a large margin. Further, using no verification method (QAEval-IsAns) largely performs equally well as QAEval variants which do use a verification step on the SummEval dataset, but not on TAC'08.

B Example BERTScore/LERC Improvements

Table 5 contains example expected answer and QA model prediction pairs for which BERTScore and LERC improve over exact match and token F_1. We see that the improvements come from better identifying when the phrases are paraphrases of each other, which sometimes involves world knowledge.

Metric	TAC'08						SummEval					
	System-Level			Summary-Level			System-Level			Summary-Level		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
BERTScore	.83	.85†	.68†	.50†	.50†	.40†	.84†	.91†	.75†	.37†	.35†	.27†
ROUGE-1	.79	.80	.60	.49†	.48†	.39†	.61	.62	.50	.28	.26	.20
ROUGE-2	.83	.87†	.67	.48†	.48†	.39†	.64	.60	.43	.23	.19	.14
ROUGE-L	.74	.77	.57	.46	.45	.36	.61	.48	.32	.21	.18	.14
ROUGE-SU4	.80	.83	.63	.49†	.48†	.39†	.62	.56	.38	.23	.19	.15
QAEval-IsAns	.87	.82	.63	.48†	.47	.37	.76	.86†	.70†	.33†	.32†	**.26**†
QAEval-EM	**.92**†	**.89**†	**.74**†	.35	.35	.29	.80†	.91†	.77†	.23	.23	.19
QAEval-F1	.90†	.86†	.68	.46	.45	.36	.82†	.91†	.77†	.30	.29	.22
QAEval-BERTScore	.90†	.85†	.68†	.49†	.48†	.38†	**.84**†	.89†	.77†	**.36**†	**.34**†	**.26**†
QAEval-LERC	.89†	.85†	.68†	**.50**†	**.49**†	**.39**†	.81†	**.93**†	**.80**†	.33†	.31†	.24†

Table 4: System- and summary-level correlations using Pearson's r, Spearman's ρ, and Kendall's τ.

Answer	Prediction	BSc	LERC
EU	European Union	0.73	0.84
a smaller leftist guerilla group	National Liberation Army	0.48	0.10
six-time Olympic gold medalist	Usain Bolt	0.34	0.35
Luis Enrique's side	Barcelona	0.40	0.18
emergency responders	paramedics	0.20	0.67
the child	toddler	0.38	0.45

Table 5: Examples where BERTScore and LERC improve over F_1 (all examples have an F_1 score of 0). Successfully classing these phrases requires paraphrasing (e.g., "the child" and "toddler") and, in some cases, world knowledge (e.g., Usain Bolt had won six gold medals when the article was written).