QuestEval: Summarization Asks for Fact-based Evaluation

Thomas Scialom[*‡], **Paul-Alexis Dray**[*], **Patrick Gallinari**[‡], **Sylvain Lamprier**[‡],
Benjamin Piwowarski[◇‡], **Jacopo Staiano**[*], **Alex Wang**[†]

[◇] CNRS, France
[‡] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[*] reciTAL, Paris, France
[†] New York University
thomas@recital.ai

Abstract

Summarization evaluation remains an open research problem: current metrics such as ROUGE are known to be limited and to correlate poorly with human judgments. To alleviate this issue, recent work has proposed evaluation metrics which rely on question answering models to assess whether a summary contains all the relevant information in its source document. Though promising, the proposed approaches have so far failed to correlate better than ROUGE with human judgments.

In this paper, we extend previous approaches and propose a unified framework, named QUESTEVAL. In contrast to established metrics such as ROUGE or BERTScore, QUESTEVAL does not require any ground-truth reference. Nonetheless, QUESTEVAL substantially improves the correlation with human judgments over four evaluation dimensions (consistency, coherence, fluency, and relevance), as shown in extensive experiments. We make code and models available.[1]

1 Introduction

The reliability of automatic evaluation metrics is an important factor for progress in artificial intelligence tasks, enabling the comparison and improvement of the proposed systems. The design of reliable metrics for natural language generation (NLG) systems is very challenging, and still an open research problem: Novikova et al. (2017) and Peyrard (2019) showed that current metrics do not correlate well with human judgments, and argued for the development of new evaluation metrics.

Among NLG tasks, summarization is one of the most difficult to evaluate automatically. For a given document, the number of possible correct outputs is much larger than for other tasks such as machine translation. Thus, when only a single reference summary is given – as is typically the case for large-scale summarization datasets, the correlation of standard automatic evaluation metrics with human judgments is low (Louis and Nenkova, 2013). Furthermore, since a summary must be shorter than the corresponding source document, information selection (Li et al., 2018) is critical so that the summary only contains the salient contents from its source document. For these reasons, n-gram based metrics, such as ROUGE (Lin, 2004), are known to poorly reflect human preference (Louis and Nenkova, 2013; Novikova et al., 2017; Paulus et al., 2017; Bhandari et al., 2020). Finally, it is crucial for an effective summarization to generate texts that are factually consistent with their source documents. However, this aspect is not measured by n-grams based metrics. Notably, while recent state-of-the-art generative models (Lewis et al., 2019; Zhang et al., 2019a) produce fluent summaries, they frequently contain false or unsupported information (Kryściński et al., 2019), a phenomenon also known as neural hallucination (Rohrbach et al., 2018; Zhao et al., 2020).

To overcome these limitations, a new approach to evaluate summarization systems has recently emerged, based on question generation (QG) and answering (QA) (Chen et al., 2017; Scialom et al., 2019; Eyal et al., 2019). These metrics measure to which extent a summary provides sufficient information to answer questions posed on its corresponding source document. They can be used to assess the factual consistency (i.e. precision) (Durmus et al., 2020; Wang et al., 2020) or the relevance (i.e. recall) (Scialom et al., 2019) of the evaluated summary, with respect to its source document. Although these works have introduced an interesting and novel method to evaluate summarization, with encouraging preliminary results, none of those metrics is found to perform better than ROUGE (Fabbri et al., 2020): automatic evaluation of summarization systems remains an open research problem (Kryscinski et al., 2019).
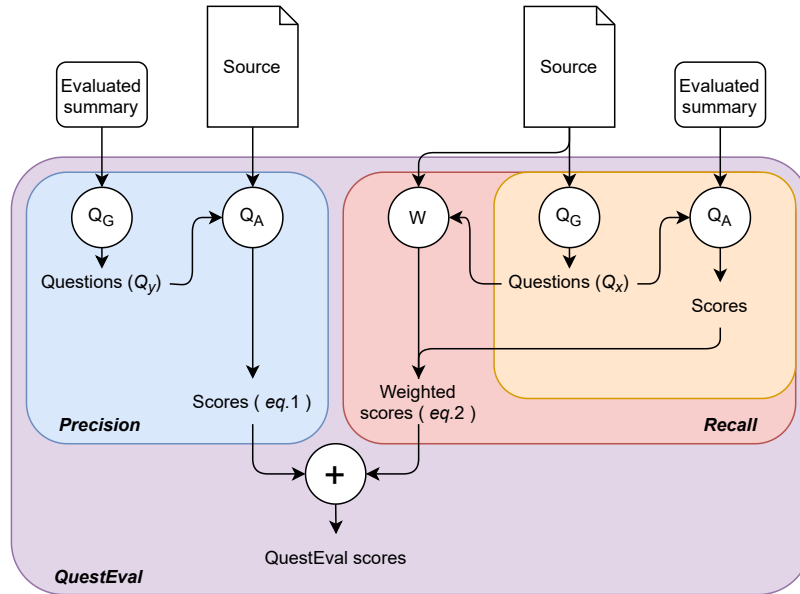
---

[1] https://github.com/ThomasScialom/QuestEval

6594

Figure 1: Illustration of the QUESTEVAL framework: the blue area corresponds to the precision-oriented framework proposed by Wang et al. (2020). The orange area corresponds to the recall-oriented SummaQA (Scialom et al., 2019). We extend it with a weighter component for an improved recall (red area). The encompassing area corresponds to our proposed unified approach, QUESTEVAL.

Inspired by these works, and motivated to take up the challenge of summarization evaluation, we propose QUESTEVAL, a new *reference-less* metric, which is found to improve the correlation with humans judgments. Our contributions are as follows:

- We show that, by unifying the precision and recall-based QA metrics, we obtain a more robust metric;

- We propose a method to learn the saliency of the generated queries, allowing to integrate the notion of information selection;

- We evaluate QUESTEVAL on two corpora containing annotated summaries from CNN/Daily Mail and XSUM datasets. The proposed metric obtains state-of-the-art results in terms of correlation with humans judgments, over all the evaluated dimensions. Notably, QUESTEVAL is effective at measuring factual consistency, a crucial yet challenging aspect for summarization.

2 Related Work

Summarization Metrics The most popular evaluation metric for summarization is ROUGE (Lin, 2004), which computes the recall of reference n-grams in the evaluated summary. Other n-grams based metrics have been proposed such as CIDEr (Vedantam et al., 2015) and METEOR (Lavie and Agarwal, 2007), but none of them correlates better with humans according to SummEval, a recent large study conducted by Fabbri et al. (2020).

Recent works have leveraged the success of pretrained language models. Zhang et al. (2019b) proposed BERTScore, which uses BERT (Devlin et al., 2018) to compute a similarity score between the reference and the evaluated text. However, its performance is similar to that of ROUGE (Fabbri et al., 2020). Several works have explored using natural language inference (NLI) models to evaluate the factual consistency of summaries (Kryściński et al., 2019; Falke et al., 2019; Maynez et al., 2020), finding mixed results in using NLI models rather than QA models.

Content Weighting Information selection is crucial in summarization: a good summary should contain only relevant facts. However, very few methods have been proposed to tackle this aspect; all the aforementioned metrics only reflect implicitly information selection, trough the gold-reference. Xu et al. (2020) proposed to explicitly model it via a function that weights the source document tokens. As opposed to our proposed weighting method, Xu et al. (2020) function requires the ground truth and thus is not suited to a reference-less scenario.

QA-Based Metrics QA-based approaches for summary evaluation were proposed a decade ago

by Clarke and Lapata (2010) for human evaluation. Chen et al. (2017) and Eyal et al. (2019) proposed to automate this approach by automatically generating questions from the reference summary. Scialom et al. (2019) extended these works by generating the questions from the source document, probing the output summary for information retrieved from the input text (a recall-oriented approach). However, by weighing each question equally, their approach lacks a way to select questions that reflect the most important information of the input.

Conversely, Wang et al. (2020) and Durmus et al. (2020) proposed to generate questions from the evaluated summary. These methods are precision oriented, since they measure the amount of information in the evaluated summary that are supported by the input text. We show in this paper that combining these recall and precision approaches leads to an improved metric.

3 A Question-Answering based Framework

This paper introduces the QUESTEVAL framework for evaluating summarization systems, which accounts for both factual consistency and relevance of the generated text, without requiring any human reference. QUESTEVAL consists of a QG component Q_G and a QA component Q_A, described in this section and depicted in Figure 1.

3.1 Question Answering

Recently, there has been significant progress on factoid question answering, with models obtaining human-level performance on benchmarks such as SQuAD (Rajpurkar et al., 2016). Leveraging on these advancements, our Q_A component consists of a pretrained T5 model (Raffel et al., 2019), which extracts answers from a source document given a question to answer. In the following, we refer to $Q_A(r|T,q)$ as the probability of the answer r to question q on a text T, and $Q_A(T,q)$ as the answer greedily generated from the model.

When a summary is evaluated, there is no guarantee that it contains the answer. Therefore, it is crucial for the QA model to be able to predict when a question is unanswerable. Our Q_A component thus includes the *unanswerable* token, that we denote ϵ, among its possible outputs.

3.2 Question Generation

For the QG component, we draw on recent work on neural answer-conditional question generation (Zhou et al., 2017). The component also consists of a T5 model, finetuned to maximize the likelihood of human questions, given the corresponding answer and source document.

At test time, given a source document or generated summary, we first select a set of answers from the text to condition the QG model on. Following Wang et al. (2020), we consider all the named entities and nouns from the source document as answers. Then, for each selected answer, we generate a question via beam search.[2]

We filter out every question for which the QA model predicts an incorrect answer. Based on this, we denote $Q_G(T)$ the set of question-answer pairs (q, r) for a text T such that $Q_A(T, q) = r$.

4 The QUESTEVAL metric

In the following, D and S are two sequences of tokens with D denoting the source document and S the corresponding evaluated summary.

4.1 Precision

A summary is deemed inconsistent w.r.t. its source text if, given a question, the answer differs when conditioned on S or D. Therefore, we define the precision score for the evaluated summary as:

$$Prec(D, S) = \frac{1}{|Q_G(S)|} \sum_{(q,r) \in Q_G(S)} F1(Q_A(D, q), r)$$

(1)

The F1 score is a standard metric for evaluating factoid question answering models, and measures the overlap between the predicted answer and the corresponding ground truth. It outputs 1 for an exact match between both answers and 0 if there is no common token. This definition of factual consistency corresponds to the frameworks proposed by Wang et al. (2020) and Durmus et al. (2020).

4.2 Recall

While a summary should contain only factual information (precision), it should also contain the most important information from its source text (recall). Extending Scialom et al. (2019) by introducing a

---

[2]We experimented with nucleus sampling (Holtzman et al., 2019) to increase diversity of the questions, with no success.

query weighter W, we define recall as:

$$Rec(D, S) = \frac{\sum\limits_{q,r \in Q_G(D)} W(q, D)(1 - Q_A(\epsilon | S, q))}{\sum\limits_{q,r \in Q_G(D)} W(q, D)} \quad (2)$$

where $Q_G(D)$ is the set of all question-answer pairs for the source text D, and $W(q, D)$ is the weight of query q for text D.

Answerability and F1 Factoid question answering models are commonly evaluated using F1 score, measuring the overlap between the predicted answer and the corresponding ground truth (Rajpurkar et al., 2016). However, an answer could be correctly expressed in different ways, e.g. "ACL" and "Association for Computational Linguistics". Unfortunately, the F1 score is 0 in this example.

To sidestep this issue, Scialom et al. (2019) use the QA confidence of answerability, i.e. $1 - Q_A(\epsilon)$, rather than F1 score. Defining recall this way allows to measure answerability independently of the way the answer is expressed, but does not take into account possible model hallucinations, i.e. the summary could answer the question incorrectly.

Conversely, when we assess factual consistency, it is not enough for a question from the summary to be answerable from the source document. The two answers to this question should *also share the same meaning* to be factually consistent. While using answerability allows for more true positives (e.g. "ACL" in the example above), for precision it is crucial to detect true negatives. This motivates our use of the F1 score in this case, similar to Wang et al. (2020).

Query Weighting In Scialom et al. (2019), all questions are considered equally important, i.e. the weight $W(q, D) = 1$ for every query $q \in Q_G(D)$. However, since a summary necessarily has a constrained length, an effective summary should contain the most important information from the source. To account for this, we introduce a question weighter, which is trained to distinguish *important* questions from *anecdotal* ones. We leverage existing summarization datasets to create training data for the weighter: given a source document D, each question $q \in Q_G(D)$ is labeled as *important* if the corresponding human summary contains the answer, as computed by the QA component applied on the summary (i.e. $Q_A(S, q) \neq \epsilon$).

$W(q, D)$ denotes the probability that q is important for D. Note that the question weighter only

concerns recall, and therefore is not applied when computing precision.

4.3 Unifying Precision and Recall

The final QUESTEVAL score accounts for both the precision and recall by computing their harmonic mean (i.e. the F-Score): $2\frac{Prec \cdot Rec}{Prec + Rec}$. The QUESTEVAL score is thus directly comparable with existing evaluation metrics, such as ROUGE or BLEU, as it lies in the same numerical range.

5 Experiments

5.1 Summarization Datasets

To evaluate QUESTEVAL, we measure its correlation with human judgments on different datasets:

SummEval Released by Fabbri et al. (2020), it is one of the largest human-annotated datasets for summarization. Derived from CNN/Daily Mail (Nallapati et al., 2016), it consists of 12,800 summary level annotations. To ensure diversity, the summaries were generated from 16 different summarization models, including extractive and abstractive architectures. To assess quality, three experts annotated four dimensions: *i) Consistency*: the proportion of facts in the summary corresponding to facts in the original text; *ii) Coherence*: how well-structured and well-organized is the summary; *iii) Fluency*: how fluent the summary is to read; and, *iv) Relevance*: the ratio between important and excess information in the summary.[3]

QAGS-XSUM Wang et al. (2020) released a subset of 239 BART outputs fine-tuned on XSUM (Narayan et al., 2018).[4] Three annotators measured the consistency of each summary.

5.2 Question Answering & Generation

To train our Q_G and Q_A models, we used the SQuAD-v2 (Rajpurkar et al., 2018) factoid question answering dataset: it is composed of (paragraph, question, answer) triplets, and includes unanswerable questions. Note that QG can be seen as the dual task for QA: any QA dataset can be used for QG by switching the generation target from the answer to the question.

Lastly, we found it helpful to train our QA model using additional synthetic unanswerable questions.

---

[3]See 4.3 Human Annotations in Fabbri et al. (2020) for more details.

[4]Note that XSUM provides more abstractive summaries than those of CNN/Daily Mail.

	#Ref	Consistency	Coherence	Fluency	Relevance	Average
ROUGE-1	11	18.1	20.1	14.9	35.6	22.2
ROUGE-L	11	15.7	15.6	13.8	33.4	19.6
METEOR	11	3.3	2.9	7.1	-0.5	3.2
BLEU	11	17.5	22.	13.7	35.6	22.2
BERTScore-f	11	20.3	18.5	21.6	31.9	23.1
ROUGE-1	1	11.0	9.8	7.5	18.9	11.8
ROUGE-L	1	8.2	7.3	5.7	13.5	8.7
BLEU	1	8.9	3.9	4.0	12.7	7.4
BERTScore-f	1	8.7	9.8	10.6	17.9	11.8
SummaQA	0	8.3	8.0	-2.9	26.2	9.9
QAGS (our impl.)	0	20.4	7.7	16.8	9.1	13.7
QUESTEVAL$_{W=uniform}$	0	43.7	22.9	28.2	37.5	33.1
w/o QA neg sampl.	0	42.5	22.5	27.7	37.2	32.4
QUESTEVAL$_{W=learned}$	0	42.0	**24.0**	28.4	**39.2**	**33.5**
Precision Only	0	**46.5**	14.0	**30.9**	22.2	28.4
Recall Only	0	30.5	22.6	19.2	37.6	27.5
NewsQA	0	40.6	22.8	27.7	38.3	**33.5**

Table 1: Summary-level Pearson correlation coefficients for various dimensions between automatic metrics and human judgments on SummEval. The top section corresponds to correlations for metrics computed on 11 reference summaries, as reported in Fabbri et al. (2020). The second section corresponds to these metrics, but given only one reference: each of the 11 available references is used alone, and the correlations averaged. The third section corresponds to the QA-based baselines. The bottom section corresponds to QUESTEVAL and its ablations.

This is done by considering a shuffled version of the dataset, where each question is randomly assigned to a paragraph from another triplet of the dataset. We consider these additional samples, with flipped contexts, as unanswerable. All experiments in this paper, except otherwise specified, use this additional data to improve identification of unanswerable queries.

5.3 Baselines Metrics

As baselines, we considered the following:

N-gram based ROUGE (Lin, 2004) is the most widely used evaluation in summarization. This metric measures the recall of reference n-grams in the evaluated summary. Conversely, BLEU (Papineni et al., 2002) computes the precision of summary n-grams in the references. METEOR (Lavie and Agarwal, 2007) is a variant that uses stemming, synonyms and paraphrastic matches.

LM based Leveraging recent progress in language modeling (LM), Zhang et al. (2019b) proposed BERTScore: for each token of the summary, the maximum cosine similarity is computed over contextualized token embeddings of the reference summary, and the overall mean is reported.

QA based SummaQA (Scialom et al., 2019) is a recall oriented metric, with questions generated from the source document. QAGS (Wang et al.,

2020) is a precision oriented metric, with questions generated from the summary.

5.4 Results

In Tables 1 and 2 we report the results for QUESTEVAL, along with several ablations. $W = uniform$ corresponds to setting all questions weights equal. Conversely, $W = learned$ corresponds to the weights learned as detailed in §4.2. We report the recall and precision component separately. Finally, for $W = learned$, we also report the results given a QA and QG component trained on NewsQA (Trischler et al., 2016), i.e. a different domain than SQuAD.

In Table 1, we observe that, amongst existing metrics, BERTScore achieves the best average Pearson correlation with human judgements (23.1), slightly above ROUGE-1 (22.2) and BLEU (22.2). These correlations are obtained when providing *no less than 11 gold references*, and averaging results. Given a single reference, all these correlations are halved. Most of the large scale datasets provide only one reference per example in their test set (e.g. CNN/Daily Mail and XSUM), a fact that highlights the importance of searching for more reference-efficient alternatives.

With regards to sample efficiency, QA-based metrics *do not require any references*. We expect Relevance to be better measured by Recall oriented metrics, and less so for Consistency. This is con-

Metric	Consistency
ROUGE-1	13.2
ROUGE-L	8.9
METEOR	10.0
BLEU	5.6
BERTScore	2.5
SummaQA	-
QAGS	17.5
QUESTEVAL$_{W=uniform}$	30.4
w/o QA neg sampl.	28.5
QUESTEVAL$_{W=learned}$	29.0
Precision Only	**32.7**
Recall Only	13.9
NewsQA	28.2

Table 2: Summary-level Pearson correlation coefficients for Consistency between various automatic metrics and human judgments on QAGS-XSUM. The top section corresponds to correlations for diverse metrics computed on one reference summary. The middle section corresponds to QA-based baselines. The bottom section corresponds to this work.

important	*answered*	Relevance Corr.
✓	✓	37.6
✓	✗	-33.5
✗	✓	-5.7

Table 3: Pearson correlation coefficients between human judgments (for Relevance) and the percentage of *important* and/or *answered* questions, on SummEval data.

firmed in the results, where SummaQA correlates better with Relevance than Consistency (26.2 vs 8.3), and vice versa for QAGS (9.1 vs 20.4). By unifying and extending the two, QUESTEVAL allows to take both dimensions into account, improving the average correlation by 18% (28.4 to 33.5).

The dimension benefiting the most from our question weighter is Relevance (+4%, from 37.5 to 39.2), indicating that our classifier learns *which questions target important information*. We discuss this aspect more in depth in §5.5.

Finally, we do not observe significant differences when using a QA and QG specifically trained on NewsQA. Compared to the other metrics, the improvement is remarkable (33.5 vs 11.8 for BERTScore), allowing better evaluations of the systems while not even requiring references.

5.5 Discussion

Reference-less One of the main limitations for the current metrics is that they require gold references to compute similarity scores. However, many possible summaries are valid for one source document. We argue that the universe of correct outputs is much larger than in other generation tasks such as machine translation. This explains why the correlations with human judgments are largely reduced when they are computed with only one reference instead of 11 (see Table 1: BERTScore-f drops from 23.1 to 11.8 in average, and other metrics likewise). Unfortunately, assuming the availability

of as many as 11 gold references is not realistic in most scenarios, due to the cost of obtaining reference summaries.

To complement Table 1, we report in Figure 2 the correlations for the best baselines as we progressively decrease the number of available gold references from 11 to 1. For all four dimensions and all the baselines, we observe that less references result in decreased correlation and increased variance. However, QUESTEVAL *does not require any reference*. Therefore, the improvement over the other metrics grows larger as the number of references used decreases. Furthermore, QUESTEVAL enables the evaluation of systems *even when no gold reference is available*.

Query Weighter There is no unique answer to the question "What makes a good summary?": it depends on the reader's point of view, which makes summarization evaluation challenging. For instance, given a contract, the seller and the buyer could be interested in different information within the same document. To instantiate the weighter W, we learn a specific dataset policy: "what kind of questions are likely answered in the CNN/Daily Mail training summaries?" This is a reasonable heuristic given that editors created the summaries following their specific policy.

To demonstrate the effectiveness of the weighter, we proceed as follows. We first consider that a question $q \in Q_G(D)$, generated on the source document, is *important* if the probability given by the query weighter is above a threshold, i.e. if $W(D, q) > 0.5$. We then say that a question is *answered* if the probability of being unanswerable is below a threshold, i.e. $Q_A(\epsilon|S, q) < 0.5$. Therefore, a question can belong to one of four folds, given the two above criteria (*important* and/or *answered*). In Table 3, we measure how the percentage of questions belonging to a specific fold correlates with the Relevance dimension for each gen-
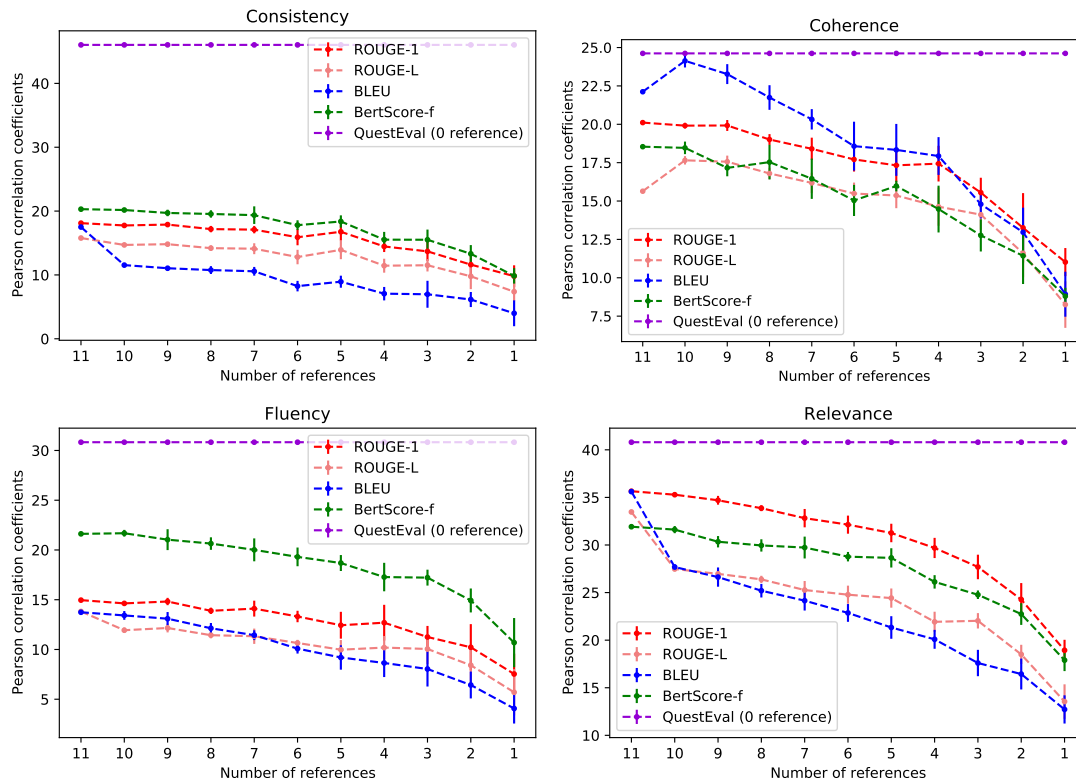
Figure 2: Variation of the Pearson correlations between various metrics and humans, versus the number of references available. QUESTEVAL is constant, since it is independent from the references.

erated summary on SummEval. We observe that the percentage of questions that are *important* and *answered* correlates positively with Relevance, as opposed to the percentage of questions that are *important* but not *answered*. Finally, the percentage of questions that are *answered* but not *important* does not correlate with Relevance. This indicates that the proposed approach is able to learn what are the questions that should be asked or not.

We emphasize that W is a flexible component of our framework: it can be adapted to specific domains and applications. For instance, one could design a specific W, to focus the evaluation on information about specific entities, such as people or events.

An Explainable Metric One important feature of QUESTEVAL is its explainability. It is straightforward to investigate *1)* what are the important points not answered in the summary and *2)* what are the inconsistencies between the source document and the summary. We illustrate this in Table 4, with a source document, from which a question q is generated and answered. According to the weighter W, q is categorized as *important*. Three evaluated summaries are then shown.

Source Document This is the embarrassing moment a *Buckingham Palace* guard slipped and fell on a manhole cover in front of hundreds of shocked tourists as he took up position in his sentry box. [...] The Guard comprises two detachments, one each for Buckingham Palace and St James's Palace, under the command of the Captain of The Queen's Guard.
Generated Question Where was the Changing of the Guard held?
Weighter prediction *Important Question*
Answer Span Buckingham Palace
Correct Summary The Queen's Guard slipped on a manhole cover during the Changing of the Guard at *Buckingham Palace* last week. [...] **Predicted Answer** Buckingham Palace: ✓
Hallucinated Summary The Queen's Guard slipped on a manhole cover during the Changing of the Guard at *St James's Palace* last week. [...] **Predicted Answer** St James's Palace: ✗
Incomplete Summary The Queen's Guard slipped on a manhole cover during the Changing of the Guard during an embarrassing moment.. [...] **Predicted Answer** Unanswerable: ✗

Table 4: Sample output from QUESTEVAL: a generated question, it's predicted importance given a source document; the corresponding predicted answers to the question, for three different summaries.
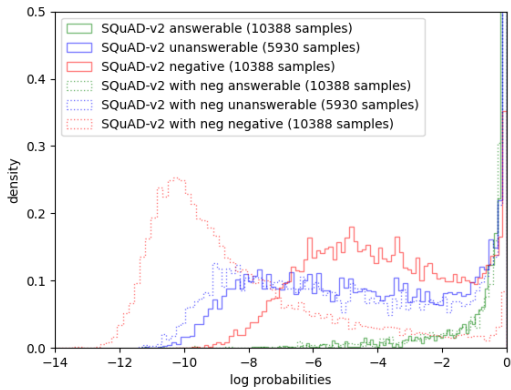
Figure 3: Distribution of the log probabilities of answerability – i.e. $\log(1 - Q_A(\epsilon|T,q))$ – for two QA models. 1) solid lines: a model trained on SQuAD-v2 *without* the negative sampled examples. 2) dashed lines: a model trained on SQuAD-v2 *with* the negative sampled examples. The evaluated samples belong to three distinct categories: 1) answerable, 2) unanswerable questions (but present in SQuAD-v2) and 3) the negatively sampled ones (as described in §5.1).

The first summary S_{correct} is factually consistent with the source document: the predicted answer $Q_A(S_{\text{correct}}, q)$ corresponds to the source document answer *Buckingham Palace*. The second summary S_{hallu} is inconsistent with the source document: the predicted answer $Q_A(S_{\text{hallu}}, q)$ *does not* correspond to *Buckingham Palace*. Finally, the third summary $S_{\text{incomplete}}$ does not answer the question, i.e. $Q_A(S_{\text{incomplete}}, q) = \epsilon$, and is thus incomplete.

Negative Sampling Effect In Tables 1 and 2, we observe a decrease of performance when QUESTEVAL uses a QA model trained without negative sampling (see Section 5.2), from 33.3 to 32.4 on SummEval and from 30.4 to 28.5 on QAGS-XSUM. In Figure 3, we report the distribution of the log probabilities for the two QA models, trained with and without negative sampling. The QA model exposed to the negative sampling during training, learns to separate better the negative sampled questions (for negative, i.e. red lines, the dashed line is more on the left than the solid line).

Indeed, the unanswerable questions of SQuAD-v2 were written adversarially by crowd-workers, to look similar to answerable ones. However, in the context of QUESTEVAL, unanswerable questions are not adversarial: it simply often happens that the summary does not contain the answer. Therefore, QUESTEVAL deals in practice with unanswerable questions that look like those built with negative

sampling, rather than adversarial ones. This may explain the improvement of a QUESTEVAL with a QA model trained with negative sampling.
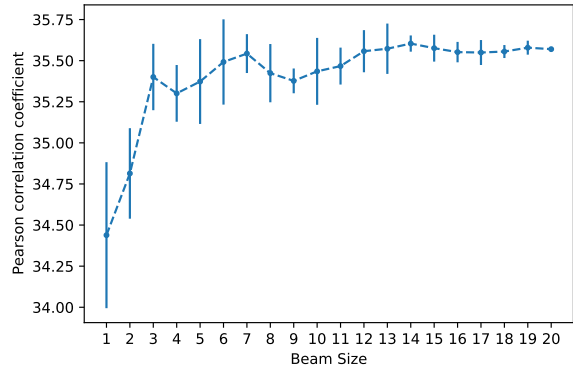

Figure 4: Pearson correlation with humans on SummEval w.r.t. the QG beam size.

Computational Complexity Following Wang et al. (2020), we generate the questions with $K = 20$ beams during decoding and we keep all the different versions of the questions in the latter steps, which improves correlations. However, the downside of this is the inference time which increases linearly w.r.t the beam size. To be widely adopted, a metric should not only correlate with human judgment, but also be computationally efficient. In Figure 4 we show the variation of the average correlation with respect to the beam size. The improvement from $K = 1$ to $K = 20$ is small (34.4 to 35.6), and the rank order for the different systems remains unchanged. Therefore, we believe that using QUESTEVAL with $K = 1$ is a reasonable choice, allowing for fast computation while preserving satisfying correlation with human judgments.

6 Conclusion

We proposed QUESTEVAL, a new *reference-less* framework to evaluate summarization models, which unifies previous QA-based approaches and extends them with question weighting, accounting all in one for factual consistency, relevance and information selection.

Compared to existing metrics, we find that QUESTEVAL correlates dramatically better with human judgments, while at the same time not requiring any reference. This allows for more accurate comparison between systems. Moreover, any progress in question answering and generation can directly be applied within our proposed framework,

leading to additional improvements. We make the code available[5] with the hope that it will contribute to further progress in the field.

We have started to adapt QUESTEVAL to other Natural Language Generation tasks that suffer from the same evaluation limitations, e.g. Text Simplification (Scialom et al., 2021), Image Captioning (Lee et al., 2021), and Data To Text (Rebuffel et al., 2021). In future work, we plan to extend QUESTEVAL to multilingual scenarios such as Machine Translation and Multilingual Summarization.

Acknowledgments

References

Manik Bhandari, Pranav Gour, Atabak Ashfaq, and Pengfei Liu. 2020. Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics. In *Proceedings of COLING 2020, the 30th International Conference on Computational Linguistics: Technical Papers*. The COLING 2020 Organizing Committee.

Ping Chen, Fei Wu, Tong Wang, and Wei Ding. 2017. A semantic qa-based approach for text summarization evaluation. *arXiv preprint arXiv:1704.06259*.

James Clarke and Mirella Lapata. 2010. Discourse constraints for document compression. *Computational Linguistics*, 36(3):411–441.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. *arXiv preprint arXiv:1906.00318*.

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

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019.

Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

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

Hwanhee Lee, Thomas Scialom, Seunghyun Yoon, Franck Dernoncourt, and Kyomin Jung. 2021. Qace: Asking questions to evaluate an image caption. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Findings*. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. 2018. Improving neural abstractive document summarization with explicit information selection modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Brussels, Belgium. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings*

---

[5]EmptyURL-Anonymoussubmission

of the 58th Annual Meeting of the Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-questeval: A referenceless metric for data to text semantic evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021. Rethinking automatic evaluation in sentence simplification. *arXiv preprint arXiv:2104.07560*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Xinnuo Xu, Ondřej Dušek, Jingyi Li, Verena Rieser, and Ioannis Konstas. 2020. Fact-based content weighting for evaluating abstractive summarisation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5071–5081.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312*.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 662–671. Springer.

Implementation details As any NLP task can be framed under the text-to-text paradigm, allowing to remove task specificities from the architecture (Raffel et al., 2019), we build QG, QA, and P, as

seq2seq models. We use the T5-base model (Raffel et al., 2019) implemented in Hugging Face (Wolf et al., 2019). We trained our models on a single Nvidia RTX 2080 Ti GPU with 11G RAM.

For selecting entities as the answers candidates, we use spaCy 2 [6].

For all the experiments, we used the default hyper-parameters.

For all the variations of hyper-parameters and models, we tested `QuestEval` on QAGS-XSUM data, in order to keep SummEval unseen at test time.

When we compute `QuestEval` precision or recall only , we observe a significant improvement over both QAGS and SummaQA. This improvement could be due to implementation differences: Scialom et al. (2019) use a rule based system to generate SummaQA's questions, and QAGS filters the questions via various heuristics. Instead, we rely on a T5-base neural generator and use only one simple heuristic. Both QAGS and SummaQA use BART/BERT for their models, while we use a smaller model, T5-base, coupled with our negative sampling method. Each of these changes improve the results, their combination showed that they add up.

Computational Complexity We believe it is important to develop effective methods before finding ways to speed them up. Despite being slower than ROUGE, QuestEval correlates much better with human judgments while not needing actual human annotators. The current running time with a single RTX 2080 is 2.53sec per document on average on CNN/DM.

Now that QuestEval effectiveness is confirmed, we plan to focus in future work on speeding up its implementation. Distilled models seems a promising direction.[7]

Moreover, a large space for improvement is possible with implementation tricks, e.g. caching the results, since always the same questions are generated. In particular, what takes most of the time is the generation of the questions on the source document: 1) it is an autoregressive process, and 2) the source document is longer than the summary, hence contains more questions. However, those questions are required to be generated only once and for all, since the source document remains un-

changed. By removing it we lower the computation to 1.82sec/doc. We plan to release an optimised version of QuestEval, including caching ability, as well as releasing all the questions on the development and test sets for XSUM & CNN/DM, which will reduce the computation cost for future evaluation.

---

[6] https://spacy.io/api/entityrecognizer
[7] https://efficientqa.github.io/assets/report.pdf