Data-QuestEval: A Referenceless Metric for Data to Text Semantic Evaluation

Clément Rebuffel [*] [1,2], **Thomas Scialom** [*] [1,3],
Laure Soulier[1], **Benjamin Piwowarski**[1], **Sylvain Lamprier**[1],
Jacopo Staiano[3], **Geoffrey Scoutheeten**[2], **Patrick Gallinari**[1,4]
[1] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[2] BNP Paribas, Paris, France
[3] reciTAL, Paris, France
[4] Criteo AI Lab, Paris, France

Abstract

In this paper, we explore how QUESTEVAL, which is a Text-vs-Text metric, can be adapted for the evaluation of Data-to-Text Generation systems. QUESTEVAL is a reference-less metric that compares the predictions directly to the structured input data by automatically asking and answering questions. Its adaptation to Data-to-Text is not straightforward as it requires multi-modal Question Generation and Answering (QG & QA) systems. To this purpose, we propose to build synthetic multi-modal corpora that enables to train multi-modal QG/QA. The resulting metric is reference-less, multi-modal; it obtains state-of-the-art correlations with human judgement on the E2E and WebNLG benchmark.[1]

1 Introduction

Automatic evaluations of NLG systems are currently based on Text-vs-Text comparison. In their most common form, these evaluations are based on n-grams comparisons between the system output and a gold reference, as with BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004). However, recent research suggests that those widely used metrics may be ill suited for NLG tasks outside of Machine Translation: BLEU, for instance, has been shown to be weak in distinguishing outputs of similar quality (Novikova et al., 2017), and dependant on contextual factors. Another issue arises in tasks with less constraints on generated texts, such as Data-to-Text Generation where diverse realisations of a single input are often acceptable, and do not necessarily share common words. To remedy this issue, a growing body of work has proposed various metrics, including a number of emerging metrics relying on specifically trained neural models (Kusner et al., 2015; Zhang et al., 2020), which compare semantic representations of texts in a latent space. However, these metrics still rely on gold references which might be biased, or even out-right incorrect or incomplete. In a context where different realisations are acceptable, even a correct gold reference is incomplete.

Recently, novel reference-less metrics have emerged to evaluate summarization abilities, relying on Question Generation (QG) and Question Answering (QA) systems (Chen et al., 2017; Scialom et al., 2019, 2021). To measure semantic matching between an evaluated summary and its source document, a QG system generates a set of relevant questions. These questions are asked on both the summary and its source document: if the answers are similar, the summary is deemed consistent with its source document.

This evaluation protocol has several benefits, most notably that it is reference-less and can fit any morphological structure of language, as opposed to n-gram based methods (Lee et al., 2020).

However, the approach proposed by Scialom et al. (2021) leverages QG and QA systems trained on dedicated corpora (e.g. SQuAD (Rajpurkar et al., 2016)). Unfortunately, the inputs in Data-to-Text Generation tasks aren't textual but rather structured data (e.g. tables) which means that existing QG/QA metrics cannot be used out-of-the-box. Dealing with this new modality requires non-trivial adaptations for existing approaches to work. More specifically, we need QG/QA models to be able to generate and answer questions given structured data.

To this end, we propose a simple yet effective method: we leverage existing QG systems trained on text inputs to create a synthetic corpus of (table, questions) pairs. By producing questions from the reference associated to each data input we are able to create a corpus of (table, question-answer) pairs, and then train QG/QA systems. QUESTEVAL can use these multi-modal models off-the-shelf, com-

[*] Equal contribution

[1] Experiments for E2E are not completed yet; results will be updated shortly.

paring directly the evaluated text to its structured data input in a reference-less scheme. We note that this method has the merit to not rely on any task-specific annotated QA dataset, which makes the approach general and suitable for any DTG task.

All in all, our contributions are three-fold:

1. We propose a methodology to train new Question Generation/Answering systems able to deal with the modality specific to Data-to-Text Generation tasks;

2. We evaluate our approach on three standard benchmarks and show that it outperforms commonly used metrics with regards to correlation with human judgement;

3. We explore practical usage of the proposed metric via a number of quantitative experiments including cross-domain adaptation.

2 Related Work

Given the limitations of BLEU, a number of approaches have been proposed to evaluate the quality of systems in Data-to-Text Generation (DTG) tasks. PARENT (Dhingra et al., 2019) improves over BLEU by also computing precision against the source table: n-grams present in the source table but omitted in the gold reference are counted as correct. Other neural metrics not specific to DTG have also been proposed, based on comparisons in the latent semantic space, such as WMD (Kusner et al., 2015) or BERTScore (Zhang et al., 2020) which compare the embeddings of output and reference according to a human-defined distance metric, or BLEURT (Sellam et al., 2020) which is trained to output similarity scores between two utterances as a regression task.

A first attempt at reference-less quality evaluation, Ranting (Dusek et al., 2019), proposed to leverage a corpus of annotated (source table, system output, human rating) triples in order to train a neural model to predict human rating as a regression task, or compare two system outputs in a "better vs worse" fashion. This approach, however, requires a significant amount of human annotated data to train the neural scorer, which is eventually biased towards the annotator rather than the task (Geva et al., 2019). Recently, Scialom et al. (2019) proposed a metric that measures the amount of information shared between two texts by generating and asking questions. It allows one to compare directly the evaluated text to its source text, without requiring any reference. Its recent extension, QUESTEVAL (Scialom et al., 2021) has shown to overperform a number of standard metrics (i.e. BLEU, ROUGE and BERTScore) in Text-vs-Text tasks such as Summarization.

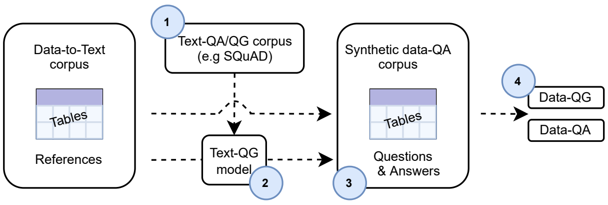

Figure 1: **Synthetic corpus creation** We are able to create a dataset of (table, question, answer) triples, by transcribing references into questions via a textual QG-model trained on SQuAD. Numerals refer to steps explained in Section 3.2.

Unfortunately, QG/QA are not usable off the shelf, as they are based on a modality (text) different from the structured data found in DTG datasets. Indeed, they are trained to produce questions/answers based on text input, for example using (context, question, answer) triples from the SQuAD dataset (Rajpurkar et al., 2016). While there have been several works on QA from structured data (e.g. the WikiTableQuestions benchmark (Pasupat and Liang, 2015)), existing approaches result in domain/task-specific systems (Angeli et al., 2010; Gatt and Krahmer, 2018) due to the high variability of datasets (different domains, different ways of structuring data, etc.) and none of the proposed data-QA datasets have significant overlaps with standard DTG benchmarks, making existing data-QG/QA systems unfit to evaluate generated outputs.

3 Our approach

3.1 QUESTEVAL for Data-To-Text

To evaluate semantic matching between two input/output texts (e.g. a document and its summary), QUESTEVAL (Scialom et al., 2021) proceeds in two steps: 1) a Question Generation system proposes (question, answer) pairs based on the input; 2) a Question Answering system proposes answers to these questions based only on the output. Semantic matching is then computed as a measure of correct answers.

To apply QUESTEVAL to DTG, and still remain in the textual modality, we first propose a simple baseline: we compare the evaluated description with the *(textual) reference*, instead of the (data) source as it was originally proposed. Because the predicted description and the reference are both texts, this approach enables us to re-use existing QG/QA models trained on a textual dataset. However, this is not satisfactory since the metric is not reference-less.

In what follows, we present our proposed method to make QUESTEVAL data-compatible, allowing it to measure the similarity directly between the source and the evaluated description.

3.2 Reference-less Multi-modal Evaluation

To make QG/QA metrics usable for reference-less evaluation on DTG tasks, specific QG/QA datasets are needed to train data-aware systems. To the best of our knowledge, there are no corpus whose domain also overlaps with traditional Data-to-Text description benchmarks. Moreover, relying on an existing corpus is not generalizable: one can not expect a data-QA dataset for every DTG dataset. The annotation for such corpora is costly and time consuming. For this reason, we propose a general approach that could be applied to any DTG dataset, and requires no specific data-QA dataset. The overall process entails four progressive steps (illustrated in Figure 1):

1) Textual QG First, following QUESTEVAL, we train a textual QG model on SQuAD.

2) Synthetic Questions Given any DTG dataset, we use the gold references as a textual input fed to our Textual QG model. We produce all the possible Questions for each example.

3) Synthetic data-QA dataset Each example is constituted of i) the source (i.e. a structured data), ii) the gold reference, and iii) the synthetic Question pairs generated in the previous step. We can therefore map to all the sources their corresponding set of synthetic Questions.

4) data-QA/QG model training Now that we have built a synthetic data-QA corpora, we can train our multi-modal QA/QG models. For QA, a source corresponds to the structured data and a synthetic Question; the target is the corresponding answer. QG can be seen as the dual task of QA; any QA dataset can be adapted into a QG dataset by considering the question as the target. To effectively tackle structured data, approaches have been proposed (e.g. a hierarchical network (Fang et al., 2019)). In this work, we adopt the T5(Kale and Rastogi, 2020) paradigm that consider any task as a Text to Text task. Therefore we simply linearize the tables and encode them directly using T5. We discuss this choice in Section 4.5.

3.3 Answer Similarity

In Question Answering, answer correctness (i.e. did the system find the correct answer?) is traditionally measured via F1-score, as popularized by the SQuAD evaluation script (Rajpurkar et al., 2016). However, F1-score is based on exact-matching of n-grams and is not accurate when evaluating a correct answer that is realised differently from the reference (e.g. a synonym). This is especially concerning in DTG, where input tables often contain data that are not found verbatim in texts (e.g. "*Place of birth: France*" can be realized as "*She is a French [...]*"). To deal with this issue, we move away from the F1-score and decide to measure answer's correctness via BERTScore, which compares the two contextualised representation of the compared answers. It allows to smooth the similarity function and provides a more accurate view of answer correctness in most DTG settings.

4 Experiments

In this paper, we want to evaluate if our proposed method to adapt QUESTEVAL in a multi-modal scenario is effective. A metric performance is measured by how much it reflects human judgement. To this purpose, we compute the correlation between human rating and QUESTEVAL, as well as several baseline metrics.

4.1 Metrics

We compare our approach to four automated metrics (2 n-gram based and 2 neural based):

BLEU (Papineni et al., 2002) computes precision of n-grams from the predicted text against the gold reference.

PARENT (Dhingra et al., 2019) is a DTG-specific metric similar to BLEU, that also includes n-grams from the source data in the computation, to benefit systems which generate true statements, even-though they are not mentioned in the gold reference.

BERTScore (Zhang et al., 2020) is a neural metric which computes a similarity score for each token in the candidate sentence with each token in the

Metric	Reference-less	WebNLG			E2E		
		Fluency	Grammar	Semantic	Fluency	Grammar	Semantic
BLEU	✗	41.0	41.8	51.4	-	-	-
PARENT	✗	47.33	50.33	63.99	-	-	-
BERTScore	✗	**58.5**	**63.8**	60.8	-	-	-
GPT-2 Perplexity	✓	49.4	56.1	48.7	-	-	-
QuesEval ref	✗	55.0	59.9	71.0	-	-	-
QuesEval src	✓	57.6	60.4	73.5	-	-	-
QuesEval ref+src	✗	57.8	61.6	**74.1**	-	-	-
QuesEval Out-of-domain	✓	57.27	61.1	71.52	-	-	-

Table 1: Comparison of Pearson coefficient of correlation w.r.t. human judgement of considered metrics against Fluency, Grammar and Semantic on the WebNLG benchmark. QuestEval out-of-domain was trained on our synthetic E2E dataset. All scores are $p < 0.05$ significant based on T-test.

reference sentence, using cosine similarity between the contextualized embeddings.

Perplexity we used GPT-2 (Radford et al., 2019), a neural Language Model trained on millions of web pages. We score a sentence using average perplexity of GPT-2 across all words.

QuestEval we report the results for QuestEval in both reference-less and reference-aware setup: `QuesEval ref` corresponds to the textual QuestEval comparing the description to its reference. `QuesEval src` corresponds to the multi-modal version comparing the description to the source. Finally, `QuesEval ref+src` corresponds to a simple average of `QuesEval ref` and `QuesEval src`. For each experiments, unless stated otherwise, we used data-QA/QG models trained on the synthetic corpus corresponding to the evaluation (e.g. synthetic WebNLG for evaluation on WebNLG.

4.2 Datasets

We evaluate our approach on the two standard DTG description benchmarks: E2E (Dušek et al., 2020) and WebNLG (Gardent et al., 2017; Castro Ferreira et al., 2020).

The **E2E** data consists of sets of key-value pairs and corresponding descriptions in the restaurant domain.

The **WebNLG** data consists of sets of RDF triple pairs and corresponding descriptions in 16 DBPedia categories (e.g., Airport, Astronaut, Building, CelestialBody, etc.). Authors of WebNLG provide a set of $2,000$ English descriptions generated by 10 different systems and annotated with Fluency, Grammar and Semantic scores (Shimorina et al., 2018), which we use to evaluate performances of automated metrics. Note that all three dimensions are evaluated on a 3-level Likert scale and are the answers to the following three questions respectively: *Rate the fluency of the text: Does the text sound fluent and natural?Rate the grammar and the spelling of the text: is the text grammatical? Does the text correctly represent the meaning in the data?*

4.3 Implementation Details

For all our experiments, we used SacreBLEU (Post, 2018) to compute the BLEU score. For PARENT (Dhingra et al., 2019), we used the original implementation that we simply optimized to run on mutli-cpus environement[2] For BERTScore, we used the original implementation[3]. The perplexity was computed with the Hugging Face implementation of GPT2-small (Wolf et al., 2019). We make QuestEval available along with the specific DTG models for reproducibility purpose.[4] All the correlations reported in this paper were computed using the SciPy python library (Virtanen et al., 2020).

4.4 Results

We report in Table 1 the correlation scores between several automatic evaluation procedures and human judgement on the WebNLG benchmark. We note that this is the first time, to the best of our knowledge, that any of the neural metrics, i.e. BERTScore and QuestEval are evaluated on DTG.

Fluency and Grammar First, we can observe that for Fluency and Grammar, neural metrics (i.e. BERTScore, QuestEval and Perplexity) dramatically improve the correlations over n-grams based metrics (i.e. BLEU and PARENT). When comparing BERTScore to QuestEval, it seems that they perform similar with a slight edge for BERTScore.

[2] https://github.com/KaijuML/parent

[3] https://github.com/Tiiiger/bert_score

[4] https://github.com/recitalAI/QuestEval/#data2text

Source: ['101_helena | discoverer | james_craig_watson', 'james_craig_watson | deathcause | peritonitis']
Reference: james craig watson , who died from peritonitis , discovered 101 helena .

Hypothesis	Questions	Predicted Answer	Score
james craero watson is the discoverts of james patson and he died in california .	Where did james patson die?	Unanswerable	0.0
james craig watson , who died of peritonitis , discovered 101 helena .	Who discovered 101 helena?	james craig watson	1.0

Table 2: We randomly sampled an example and the prediction of two different systems and report for each of them an example of an automated generated question and its predicted answer.

Semantic Regarding the Semantic dimension, we note that BERTScore correlation is lower than PARENT, while QuestEval obtains a large improvement (74.1 vs previous best of 64 by PARENT). We stress that Semantic is one of the most important dimension to measure: current models have shown to be fluent but hallucinating. These results indicate that QuestEval can be efficiently adapted to evaluate DTG.

QuestEval: using the reference or not? In our ablation studies, we compare the effect of using or not the gold reference in QuestEval. We can observe that using only the source performs even better than using only the reference. We hypothesise that some references fail to give accurate descriptions of the tables, which might explain the lower performance. This emphasizes the interest for an evaluation metric able to compare the evaluated text directly against the source. Nonetheless, the correlations benefit from using both the source and the reference, achieving the best performance (i.e. 74.1 for `QuestEval ref+src`).

4.5 Discussion

This work consolidates results from (Scialom et al., 2019, 2021) on QG/QA-based metrics for the automated evaluation of generated texts. In the process of adapting this class of metrics to Data-to-Text Generations, we have also added a number of incremental changes which are worth discussing here.

On the QA potential QuestEval directly depends on the QA model performances. While not the focus of this study, recent works (Chen et al., 2020; Nan et al., 2021) on data-QA have shown great promise and could lead to further improvement for data-QuestEval. In a larger view, research in Question Answering has been very prolific, and further improvements, either on tables or texts, will undoubtedly lead to improvements in the quality of QuestEval's evaluation. We note that in some way, the problematic of the evaluation is deported from the metric to the QA field which is a better defined task.

On cross-domain adaptation How would QuestEval perform on some dataset, e.g. WebNLG, if its data-QA/QG components were trained on another synthetic dataset, e.g. E2E? In Table 1, we report this experiment (see `QuesEval Out-of-domain`). The results are only slightly lower that in-domain. This indicates that our approach allows to train data-QA/QG models able to generalize on other domains. We note that both E2E and WebNLG no not differ too much in the way their source tables are structured. We hypothesise that a larger drop of performance might occur otherwise, and recommend to build synthetic datasets and train specific models in this case.

An interpretable metric The way QuestEval evaluates descriptions is directly related to the answers for a set of questions. In Table 2 we provide an example of tabular data with two different evaluated descriptions, generated questions, and the predicted answers. The first hypothesis contain an hallucination: *California* was never mentioned in the source table. In accordance, our QA model predicted *Unanswerable*. This emphasizes an interesting aspect: QuestEval is explainable.

A sequence-level evaluation The DTG community is progressively progressing toward more complex tables, hence longer descriptions. In this context, the community will need metrics able to evaluate long sequences. As in a DTG task several realization of a correct description are possible, the number of potentially correct gold-references exponentially raises w.r.t. the length of the sequence. In this context, QuestEval becomes particularly interesting, as it loosens the dependence on a specific realization of the source table. As opposed to tokenlevel metrics, QuestEval is robust to sentence splitting, word reordering, and now synonyms (see Section 3.3. Moreover, QuestEval is sensible to Fluency given that it uses neural QA/QG systems based on pre-trained Language Models.

5 Conclusion

In this work we explore the evaluation of Data-to-Text Generation systems using Question Generation/Answering, namely QuestEval. We propose a methodology to create synthetic corpora of (data, question, answer) triples. The effectiveness of our proposed method enables researcher to use QuestEval for Data-to-Text tasks in a reference-less setup. We show that our approach outperforms a number of varied existing metrics w.r.t. human judgement. In future work, we plan to study how QA systems perform on more complex and abstractive datasets.

References

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–512, Cambridge, MA. Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

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

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. 2020. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *arXiv preprint arXiv:2004.07347*.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Ondrej Dusek, Karin Sevegnani, Ioannis Konstas, and Verena Rieser. 2019. Automatic quality estimation for natural language generation: Ranting (jointly rating and ranking). In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 369–376. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2019. Hierarchical graph network for multi-hop question answering. *arXiv preprint arXiv:1911.03631*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org.

Dongyub Lee, Myeongcheol Shin, Taesun Whang, Seungwoo Cho, Byeongil Ko, Daniel Lee, Eunggyun Kim, and Jaechoon Jo. 2020. Reference and document aware semantic evaluation methods for korean language summarization. In *Proceedings of COLING 2020, the 30th International Conference on Computational Linguistics: Technical Papers*. The COLING 2020 Organizing Committee.

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.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Nick Schoelkopf, Riley Kong, Xiangru Tang, et al. 2021. Fetaqa: Free-form table question answering. *arXiv preprint arXiv:2104.00369*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

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

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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*, pages 2383–2392.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Anastasia Shimorina, Claire Gardent, Shashi Narayan, and Laura Perez-Beltrachini. 2018. WebNLG Challenge: Human Evaluation Results. Technical report, Loria & Inria Grand Est.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

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 preprint arXiv:1910.03771*.

Tianyi Zhang, Varsha Kishore, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.