Data-QuestEval: A Reference-less Metric for Data-to-Text Semantic Evaluation

Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski,
Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, Patrick Gallinari

▶ **To cite this version:**

Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, et al.. Data-QuestEval: A Reference-less Metric for Data-to-Text Semantic Evaluation. 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Nov 2021, Punta Cana, Dominican Republic. pp.8029-8036, 10.18653/v1/2021.emnlp-main.633 . hal-03479905

HAL Id: hal-03479905
https://hal.sorbonne-universite.fr/hal-03479905

Submitted on 14 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:2104.07555v3 [cs.CL] 7 Sep 2021

Data-QuestEval: A Reference-less 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

QUESTEVAL is a reference-less metric used in text-to-text tasks, that compares the generated summaries directly to the source text, by automatically asking and answering questions. Its adaptation to Data-to-Text tasks is not straightforward as it requires multimodal Question Generation and Answering systems on the considered tasks, which are seldom available. To this purpose, we propose a method to build synthetic multimodal corpora enabling to train multimodal components for a data-QuestEval metric. The resulting metric is reference-less and multimodal; it obtains state-of-the-art correlations with human judgment on the WebNLG and WikiBio benchmarks. We make data-QUESTEVAL's code and models available for reproducibility purpose, as part of the QUESTEVAL project.[1]

1 Introduction

Data-to-Text Generation (DTG) aims at generating descriptions in natural language given a structured input, e.g. a table (Gatt and Krahmer, 2018). Reliability and precision of generated texts is currently regarded as a major issue in DTG (Narayan and Gardent, 2020), with experimental surveys showing that real-life end users of DTG systems care more about accuracy than about readability (Reiter and Belz, 2009). Neural NLG systems are known to be fluent, but prone to hallucinations (Lee et al., 2018), i.e. they tend to include nonfactual information. However, their evaluation remains an open research problem (Novikova et al., 2017).

A recent approach, QuestEval (Scialom et al., 2021) has shown significant improvement over standard metrics on Summarization tasks. To measure semantic matching between an evaluated summary and its source document, QuestEval relies on Question Generation and Answering (QG/QA) systems.

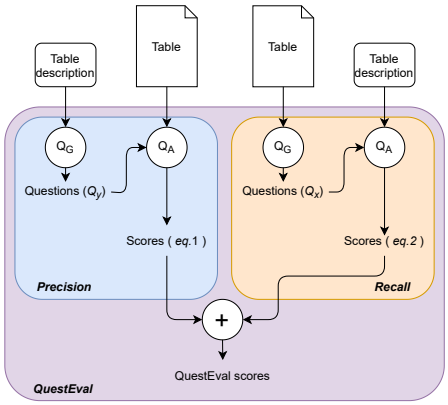

Figure 1: **Data-QUESTEVAL Flowchart.** Figure adapted from the work of Scialom et al. (2021) (equation numbers refer to equations in the original paper).

As illustrated in Figure 1, a QG system generates a set of relevant questions conditioned on the source document, which are then asked on its generated summary. Conversely, questions generated from the summary are answered using only the source input. If the answers provided by the QA systems are correct, the summary is deemed consistent with its source document.

Can QuestEval be adapted for evaluation on DTG tasks? So far, QuestEval's QG/QA systems have been trained on a purely textual dataset, SQuAD (Rajpurkar et al., 2016), which restricts the evaluation to comparisons between two texts. Unfortunately, DTG inputs are of different modalities than text (e.g. structured tables). In the absence of specific multimodal-QA datasets, how can one obtain these multimodal-QG/QA models required for a data-QuestEval?

To fill this gap, we propose an effective method for creating synthetic multimodal-QG/QA datasets, by relying only on existing, purely textual, QG/QA datasets. Trained on such synthetic multimodal datasets, QA and QG models can now be used in QuestEval, enabling direct comparison between an evaluated text and its structured input, removing the need for costly gold references. Furthermore, this

*Equal contribution
[1] https://github.com/ThomasScialom/QuestEval

method does not rely on any task-specific annotated QA dataset, which makes the approach general and suitable for any DTG task.

2 Related Work

Based on n-grams similarity between an evaluated text and its gold references, BLEU (Papineni et al., 2002) and PARENT (Dhingra et al., 2019) are the two standard metrics reported in DTG evaluations. Beyond n-grams, Opitz and Frank (2020) proposed to use a model trained on the reverse task, i.e. text to data reconstruction, and to compare the two data generated i) from the reference, and ii) from the hypothesis. Zhang et al. (2020) introduced BERTScore, where texts are compared given the contextual BERT representations of their tokens. However, all these metrics require gold references. In a first attempt for a reference-less evaluation metric in DTG, Dusek et al. (2019) proposed to train a neural model to directly predict the human ratings; this requires a significant amount of human annotated data, and is eventually biased towards the annotator rather than the task (Geva et al., 2019).

Concurrently, a family of reference-less metrics has emerged in Summarization (Chen et al., 2017; Scialom et al., 2019). The amount of information shared between two texts is measured by generating questions on the source text, and asking them on the evaluated text. In its recent extension, QUESTEVAL (Scialom et al., 2021) was shown to overperform standard metrics in Text-vs-Text tasks such as the evaluation of summaries.

Unfortunately, text-QG/QA systems are not usable off the shelf for tasks in other modalities which require multimodal-QG/QA systems. Further, there is significant variability in structures (e.g. tables, knowledge graphs, etc.) and domains (e.g. biographies, sports, etc.) across DTG tasks (Gatt and Krahmer, 2018). Therefore, generalizing QUESTEVAL to DTG tasks relying solely on existing multimodal-QG/QA datasets is not a promising direction: the effectiveness of data-QG/QA models would be limited to the specific structures of the few existing data-QA datasets – e.g. the WikiTableQuestions benchmark (Pasupat and Liang, 2015); moreover, it is unrealistic to annotate QG/QA corpora for each specific modality/domain.

Conversely, our proposed method generalises to any structured data, given only a text-QA dataset.

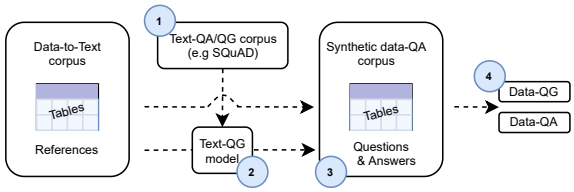

Figure 2: **Synthetic corpus creation** We are able to create a dataset of (table, question, answer) triples, by transcribing references into questions via a textual QG-model trained on SQuAD. Numerals refer to steps explained in Section 3.2.

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 generates a set of questions and (true) answers given the input text; 2) a Question Answering system predicts the (candidate) answers to these questions relying only on the output text currently evaluated. Candidate answers are evaluated based on F1-score against the true answers and Semantic Matching is then computed as the mean of all F1-scores.

To apply QUESTEVAL to DTG, and still remain in the textual modality, one can consider a simple baseline: comparing an evaluated description with its (textual) reference, instead of its (data) source. Since the predicted description and the reference are both texts, this approach enables us to re-use QUESTEVAL in its vanilla form without any multimodal requirements. However, this is not satisfactory, as this metric ignores the structured input, contrary to the original intent of QUESTEVAL. Further, this makes the metric dependent on human annotations which may be costly to obtain.

In the following, we present our proposed method to make QUESTEVAL data-compatible, allowing it to measure the similarity directly between a structured input and the evaluated description.

3.2 Reference-less multimodal 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. Relying on an existing corpus is not generalizable: it is unreasonable to expect a multimodal-QA dataset for every DTG dataset. The annotation necessary to build such corpora is costly and time consuming. For this reason, we propose a general approach applicable to any DTG dataset, and requiring no

Metric	Reference-less	WebNLG			WikiBio	
		Fluency	Grammar	Semantic	Fluency	Semantic
BLEU	✗	41.0	41.8	51.4	0.8	8.1
PARENT	✗	47.33	50.33	63.99	-1.1	9.5
BERTScore	✗	**58.5**	**63.8**	60.8	11.5	8.1
GPT-2 Perplexity	✓	49.4	56.1	48.7	7.5	7.0
text-QuestEval	✗	55.0	59.9	71.0	11.0	15.2
data-QuestEval	✓	57.6	60.4	**73.5**	**13.2**	**18.2***
OOD-data-QuestEval	✓	57.27	61.1	71.52	**13.2**	15.8

Table 1: Pearson correlation coefficients of considered metrics against human judgements of Fluency, Grammar and Semantic. On WebNLG, all scores have a p-value $p < 0.05$; On WikiBIO, * indicates $p < 0.05$.

annotated multimodal-QA dataset. The overall process entails four steps (illustrated in Figure 2):

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

Step 2) Synthetic Questions Given the training set of any DTG dataset, composed of (structured-input, textual description) pairs, we generate synthetic questions for each *textual description* using the *textual QG* (from step 1).

Step 3) Synthetic multimodal-Question dataset Each example in the training set is constituted of i) the source (i.e. structured data), ii) the textual description, and iii) the synthetic (Question, Answer) pairs generated during step 2. We can therefore match each structured input to its corresponding set of synthetic Questions & Answers to build a data-QG/QA dataset.

Step 4) Multimodal-QG/QA model training The newly built synthetic multimodal-Question corpus is used to train multimodal QG/QA 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 used as a QG dataset by considering the question as the target. To learn representations from structured data, several approaches have been proposed – e.g. a hierarchical network (Fang et al., 2019). We adopt the T5 (Kale and Rastogi, 2020) paradigm, where any task is considered as a Text-to-Text task: we linearize the tables and encode them directly using T5.

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 realized 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 (Zhang et al., 2020), which compares the two contextualized 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.

3.4 Reproducibility

Beyond enabling the comparison of different versions of a given model for a specific project, evaluation metrics make it possible to compare different models altogether between different projects. To avoid inconsistencies in the reporting of QUESTEVAL scores, our code follows the guidelines of (Post, 2018) and produces a short version string that facilitates cross-paper comparisons, identifying the model checkpoints used, preprocessing steps, etc. Reporting this version string will ensure fair comparison across future works.

4 Experiments

In this paper, we aim to evaluate the effectiveness of our proposed multimodal adaptation of QuestEval. Consistently with previous works (Dhingra et al., 2019; Zhang et al., 2020), metric performance is measured by how much it reflects human judgement, assessed via Pearson correlation coefficient.

4.1 Metrics

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

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

Hypothesis	Generated Questions	Predicted Answers	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: Example of a structured input and predictions of two different systems, along with automatically generated questions and predicted answers. *the QA system can warn that no answer can be found.

BLEU (Papineni et al., 2002) compares n-grams between the output and the reference.

PARENT (Dhingra et al., 2019) is a DTG-specific metric similar to BLEU, that also includes n-grams from the source data, to favour systems that generate true statements, which may not be 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 reference sentence, using cosine similarity between the contextualized embeddings.

Perplexity A sentence is scored using the average perplexity of GPT-2 (Radford et al., 2019) across all words.

QUESTEVAL `text-QuestEval` corresponds to the baseline presented in Section 3.1, using textual QUESTEVAL to compare evaluated descriptions to their references. In contrast, `data-QuestEval` corresponds to the multimodal version we propose in Section 3.2, comparing the evaluated description to its structured source. For all experiments, unless stated otherwise, we used multimodal-QG/QA models trained on the synthetic corpus corresponding to the evaluation data (e.g. synthetic WebNLG for evaluation on WebNLG). Finally, `OOD-data-QuestEval` was trained on a synthetic E2E dataset (Dušek et al., 2020) to assess the impact of out-of-domain training.

4.2 Datasets

We evaluate our metric on WebNLG (Shimorina et al., 2018) and WikiBio (Lebret et al., 2016).

The WebNLG data consists of sets of RDF triples and corresponding descriptions in 16 DBPedia categories (e.g., Airport, or Food). Authors of WebNLG provided a set of $2,000$ English descriptions generated by 10 different systems and rated by human annotators on a 3-level Likert scale on three dimensions: ***Fluency*** - *does the text sound fluent and natural?*, ***Grammar*** - *is the text grammatically correct?*, and ***Semantic*** - *does the text correctly represent the meaning in the data?*

The WikiBio data consists of biographies paired with the corresponding infoboxes extracted from Wikipedia biography articles. WikiBio tables are one order of magnitude larger than tables in WebNLG (see Table 3 in supplementary material), and training instances have been built automatically from online sources, resulting in very noisy reference texts (Dhingra et al., 2019; Rebuffel et al., 2021). This increased complexity shines an interesting light on metrics' performances. We used the human evaluation from Rebuffel et al. (2021), who collected ratings of three models (one baseline LSTM, and two SOTA systems) on 200 examples of WikiBio, following the protocol of Shimorina et al. (2018) and evaluating two dimensions: Fluency and Semantic.

4.3 Results and Discussion

In Table 1, we report the correlation scores between several metrics and human judgment on the WebNLG and WikiBio benchmarks.

To the best of our knowledge, this is the first work that evaluates neural metrics (BERTScore, GPT2-Perplexity and QUESTEVAL) for DTG.

WebNLG For Fluency and Grammar, neural metrics (i.e. BERTScore, QUESTEVAL and Perplexity) dramatically improve the correlations over ngram-based metrics (i.e. BLEU and PARENT). When comparing BERTScore to QUESTEVAL, while the former require a reference as opposed to the latter, the performance is almost comparable. For the Semantic aspect, we find that BERTScore correlates less than PARENT, while QuestEval shows a large improvement over the previous best score of PARENT (73.5 vs 64 resp.).

WikiBIO The WikiBIO annotation corpus differs from WebNLG in two ways: i) the input tables

are more complex (see Table 3 in supplementary material); ii) the systems used for the evaluation of WikiBIO are very recent; their fluency is close to human level – see Table 4 in (Rebuffel et al., 2021). This can explain why no metric exhibits a significant correlation for Fluency. On Semantic, the only metric that correlates significantly is QUESTEVAL, indicating the effectiveness of our proposed method to evaluate DTG. We stress that Semantic is one of the most important dimensions to measure: current models have shown to be fluent but hallucinating (Lee et al., 2018).

Finally, we observe on both datasets that QUESTEVAL performs better using the source than the reference. We hypothesize that some references fail to give accurate descriptions of the tables, which might explain the lower performance (more details in Appendix D of the supplementary materials).

On the QG/QA potential QUESTEVAL directly depends on the performances of its QG/QA models. While not the focus of this study, recent works (Chen et al., 2020; Nan et al., 2021) on multimodal-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 evaluating text is removed from the Data-To-Text task, and moved to the QA field, where it is arguably better defined.

On cross-domain adaptation How would QUESTEVAL perform if its multimodal-QG/QA components were to be trained on another synthetic dataset, such as E2E (Dušek et al., 2020)? In Table 1, we observe that the results of `OOD-data-QuestEval` on WebNLG are only slightly lower than those of QUESTEVAL in-domain, indicating that our approach generalize on similar domains. In contrast, results on WikiBio are not conclusive, highlighting that significant variations on structure and domain leads to dramatic decrease of performances across tasks.

An interpretable metric As noted by Rebuffel et al. (2021), the commonly used metrics compared in this work provide sentence-level information, which can be hard to decipher: for instance, given a 0.5 PARENT score, which part of the prediction is incorrect? In contrast, QUESTEVAL scores are directly related to its QA's answers for a set of questions generated by its QG. As such, it provides fine-grained, interpretable explanations of its score, via the analysis of the QG's questions and QA's answers (or lack thereof). Table 2 showcases an example of tabular data with two evaluated descriptions, generated questions, and the predicted answers. This emphasizes the explainability of QUESTEVAL, an interesting property that we plan to further explore in future work.

On Multilingual extensions Experiments of this paper are performed on English corpora. However, in the light of successful QG/QA approaches in different languages, we believe that our work will generalize well to other languages. In particular, languages in which ngram-based metrics perform poorly due to the language's structure (e.g. Korean (Lee et al., 2020)) would benefit the most from our approach).

5 Conclusion

In this work, we proposed an efficient method to evaluate in a reference-less multimodal setup of Data-to-Text Generation. To train Data-QuestEval on different modalities, we propose a methodology to create synthetic corpora of (data, question, answer) triples from any Data-to-Text Generation task. We show that our approach outperforms several existing metrics w.r.t. human judgement on two standard Data-to-Text Generation benchmarks, and performs reasonably well when trained on a different domain.

While QG/QA modeling per se is out of the scope of this paper, further improvements could be obtained thanks to advances in multimodal QA, a field particularly prolific (Chen et al., 2020; Herzig et al., 2020; Nan et al., 2021). In future works, we plan to study how QG/QA systems perform on more complex and abstractive datasets.

References

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*.

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.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. 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.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Dongyub Lee, Myeong Cheol 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 the 28th International Conference on Computational Linguistics*, pages 5604–5616, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

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*.

S. Narayan and C. Gardent. 2020. *Deep Learning Approaches to Text Production*.

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.

Juri Opitz and Anette Frank. 2020. Towards a decomposable metric for explainable evaluation of text generation from amr. *arXiv preprint arXiv:2008.08896*.

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.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2021. Controlling hallucinations at word level in data-to-text generation. *arXiv preprint arXiv:2102.02810*.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *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.

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*.

A Datasets

		E2E	WebNLG	WikiBIO
table size	max	8	11	86
	mean	5.37	4.5	12.42
target size	max	67	445	6340
	mean	19.72	117.08	97.02

Table 3: Lengths of inputs and outputs in E2E, WebnNLG and WikiBIO.

B 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).

C 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 realizations 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 token-level 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.

D QuestEval: using the reference or not?

In our ablation studies, Table 1 of the Main Paper, 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.

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

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

[4] *Anonymous EMNLP submission - hidden URL*