A Case for Better Evaluation Standards in NLG

Sebastian Gehrmann **Elizabeth Clark** **Thibault Sellam**

Google Research
New York, NY
{gehrmann,eaclark,tsellam}@google.com

Abstract

Evaluating natural language generation (NLG) models has become a popular and active field of study, which has led to the release of novel datasets, automatic metrics, and human evaluation methods. Yet, newly established best practices are often not adopted. Moreover, the research process is often hindered by the scarcity of released resources like model outputs, and a lack of documentation of evaluation parameters often complicates judging new NLG methods. We analyze 66 papers published in 2021 across 29 different dimensions to quantify this effect, and identify promising ways for the research community to improve reporting and reviewing experimental results.

1 Introduction

For authors and reviewers in empirical machine learning, evaluation is key to verify the validity of a scientific claim. Yet collecting and reporting experimental results involves decisions that are not always reported, including the selection of datasets, the choice and parametrization of metrics, the task setups used for human evaluation, and many more. Often these design decisions are based on previously published work and assumptions about what the community considers acceptable, a cycle that can lead to the normalization of flawed evaluation and reporting practices.

This is a particularly salient problem for text generation, where researchers have long called out opaque evaluation practices [e.g., Stent et al., 2005, Belz and Gatt, 2008, Pitler et al., 2010]. The issues range from metrics that do not correlate well with different surface-level [Fabbri et al., 2021] or semantic [Maynez et al., 2020] quality aspects, to a focus on English datasets [Joshi et al., 2020], or underreported human evaluation details [Howcroft et al., 2020]. As a result of these and many other issues, it is challenging to improve evaluations as a whole. New evaluation techniques are rarely widely adopted, and reviewers may not even be aware what constitutes a "good" evaluation.

This paper presents a list of 29 suggestions across 8 high-level evaluation aspects that is grounded in the published NLG evaluation literature, and we quantify the extent to which authors follow them. We survey 66 papers published at EMNLP, INLG, and ACL in 2021, find that these practices are followed at an average rate of 27% and uncover dimensions that require more drastic changes in the NLG community. For an extended motivation for each suggestion, we point to the extended version of this work [Gehrmann et al., 2022].

2 Background and Study Setup

Our analysis focuses on *conditional* natural language generation. We consider tasks in which a model can be trained to maximize a conditional probability $p(y|x)$ where y is natural language and x is an input that can be structured data or natural language and which provides information about what should be generated. We require the NLG tasks to have an explicit *communicative goal*, which

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

needs to be expressed. That means that a model has to plan the content and structure of the text, and actualize it in fluent and error-free language [Gehrmann, 2020]. This includes, e.g., summarization, machine translation, and paraphrasing, while excluding question answering (answer-spans are not natural language) and open ended language modeling (unconditional). We also omit multimodal tasks (e.g., image captioning, speech-to-text), as well as those with non-textual output (e.g., sign language, audio) because they require different evaluation processes.

Starting with the accepted long papers from EMNLP (848 papers), ACL (572), and INLG (46) 2021, we filtered to papers with titles that directly mentioned an NLG task or used related keywords (like "generating" or "realizing"). Papers were excluded from the final list when, upon reading them, we noticed that they either did not report any results or only for tasks not covered by the above definition. The final list includes 66 papers.

3 Evaluation Best Practices and Annotation Instructions

We annotate each paper for 29 dimensions of NLG model evaluation based on 8 categories of best practices introduced by Gehrmann et al. [2022]. The annotation instructions and limitations are in Appendix A and B, respectively. Below, we provide brief context for these suggestions, which are listed in Table 1.

Make informed evaluation choices and document them. Prior work has called out issues in the documentation of details of the ML pipeline e.g., in datasets [Bender and Friedman, 2018, Gebru et al., 2021, McMillan-Major et al., 2021], models [Mitchell et al., 2019], and human evaluations [Shimorina and Belz, 2021]. A similar argument can be made for evaluation, where, for example, the design of data splits [Søgaard et al., 2021] and the reference style [Goel et al., 2021] may favor systems by design, yet those choices are not always documented. Liao et al. [2021] point out that equating a benchmark task with insights into model capabilities can lead to harmful over-generalization. We further aim to measure the adoption of non-English datasets [e.g., Scialom et al., 2020, Ladhak et al., 2020, Hasan et al., 2021].

Measure specific generation effects. The exponential output space in NLG sets it apart from other NLP tasks and leads to a reliance on automatic metrics. However, that means that evaluation results are only as trustworthy as the metrics. Unfortunately, most commonly used metrics have a poor correlation with human judgments [e.g., Fabbri et al., 2021, Deutsch et al., 2021]. Moreover, different quality aspects (e.g., grammaticality, faithfulness) may not correlate with each other [Pitler et al., 2010, Graham, 2015, Deutsch and Roth, 2021], which suggests that a single number, as produced by almost all automatic metrics, cannot fully characterize an NLG system. Another conceptual flaw is that metrics by design are unidirectional: an increase suggests that a system is "better", but often an improvement on one axis comes at a cost in other areas. Evaluations should thus also identify these trade-offs and potential shortcomings.

Analyze and address issues in the used dataset(s). Model limitations often stem from issues in the data, and the data itself can lead to false downstream claims. To address these issues, we need to improve how data collection processes are documented [Bender and Friedman, 2018, Gebru et al., 2021]. Additionally, paying closer attention to datasets can lead to improvements for the whole research field [e.g., Dušek et al., 2019, Thomson and Reiter, 2020] over time. Sending pull requests to update data documentation and datasets thus needs to become as commonplace as sending pull requests to or opening issues in open-source libraries. Treating datasets as dynamic encourages the development of evaluation suites that everyone can benefit from [Bowman and Dahl, 2021].

Evaluate in a comparable setting. Another commonly found issue is the lack of reproducibility of evaluation numbers. Metrics have many hyperparameters and few of them are commonly reported, leading to unfair comparisons [Liao et al., 2021, Post, 2018]. Numbers should thus be recomputed in the same environment.

Run a well-documented human evaluation. Howcroft et al. [2020] find that parameters of human evaluations are often underreported, which can lead to implicit overclaims, a lack of reproducibility, and the absence of robust evaluation standards. Many aspects that should be reported are proposed in human evaluation datasheets [Shimorina and Belz, 2021, Belz et al., 2021].

Produce robust human evaluation results. In addition to better documentation, we also need to improve how human evaluations work toward reusability and replicability in human evaluations, e.g.,

Best Practice & Implementation	Yes	No	%
Make informed evaluation choices and document them			
Evaluate on multiple datasets	47	9	83.9
Motivate dataset choice(s)	21	34	38.2
Motivate metric choice(s)	20	46	30.3
Evaluate on non-English language	19	47	28.8
Measure specific generation effects			
Use a combination of metrics from at least two different categories	36	27	57.1
Avoid claims about overall "quality"	34	31	52.3
Discuss limitations of using the proposed method	19	46	29.2
Analyze and address issues in the used dataset(s)			
Discuss or identify issues with the data	19	47	28.8
Contribute to the data documentation or create it if it does not yet exist	1	58	1.7
Address these issues and release an updated version	3	10	23.1
Create targeted evaluation suite(s)	14	52	21.2
Release evaluation suite or analysis script	3	63	4.5
Evaluate in a comparable setting			
Re-train or -implement most appropriate baselines	40	19	67.8
Re-compute evaluation metrics in a consistent framework	38	22	63.3
Run a well-documented human evaluation			
Run a human evaluation to measure important quality aspects	48	18	72.7
Document the study setup (questions, measurement instruments, etc.)	40	9	81.6
Document who is participating in the study	28	20	58.3
Produce robust human evaluation results			
Estimate the effect size and conduct a power analysis	0	48	0.0
Run significance test(s) on the results	12	36	25.0
Conduct an analysis of result validity (agreement, comparison to gold ratings)	19	29	39.6
Discuss the required rater qualification and background	10	38	20.8
Document results in model cards			
Report disaggregated results for subpopulations	13	53	19.7
Evaluate on non-i.i.d. test set(s)	14	52	21.2
Analyze the causal effect of modeling choices on outputs with specific properties	16	50	24.2
Conduct an error analysis and/or demonstrate failures of a model	15	51	22.7
Release model outputs and annotations			
Release outputs on the validation set	1	65	1.5
Release outputs on the test set	2	63	3.1
Release outputs for non-English dataset(s)	1	25	3.8
Release human evaluation annotations	1	47	2.1

Table 1: Suggested best practices and number of papers that follow them. See Appendix A for exact annotation instructions.

by using projects that standardize parts of the process [e.g., Khashabi et al., 2021, Gehrmann et al., 2021]. To that regard, we measure adherence to some of the best practices suggested by van der Lee et al. [2019], effect size estimates, power analyses, statistical significance tests, and analyses of the validity of human evaluation results.

Document results in model cards. Mitchell et al. [2019] describe the "quantitative analysis" process of reporting disaggregated results according to chosen metrics. Generalizing this argument, we need to identify what breaks a model, with the goal of moving away from chasing the highest overall number. The long-term goal of evaluation reports are performance guarantees: we would like to know exactly what to expect of a model for a given input. Evaluation reports should further include improved error analyses, following suggestions by van Miltenburg et al. [2021] and Bender and Koller [2020], who argue for more focus on limitations in addition to aggregated scores.

Release model outputs and annotations. Finally, to improve replicability, model outputs for validation and test sets alongside instructions on how to replicate reported numbers should be released. Many works like that of Fabbri et al. [2021] would not be possible without access to model outputs, and such corpora can be used for metric development and validation, and to conduct meta evaluations. Releasing outputs on non-English datasets, even when no human evaluation can be conducted, supports evaluation improvements on the covered languages by reducing the burden on the evaluation researchers to produce the outputs.

4 Results

We find that 36.7% of our 2046 judgments were positive, which means that the field has already taken a significant step toward solving the problems pointed out throughout this survey. Scores for papers ranged from 6.5% to 58.1%, with an average of 27.3% (median 25.8%, standard deviation of 0.11), suggesting that no consistent standard is widely applied.

The vast majority of papers include evaluation results from multiple datasets (84%) and report human evaluation results (73%). However, the documentation of the choices that went into the evaluation process is often flawed. Only 38 and 30% of papers respectively motivate why they chose a particular dataset and metric, and half the papers made claims in the abstract pertaining to their system outputs' overall quality when this was not the aspect that was evaluated. About 29% of papers reported results on a non-English language, although most were machine translation papers. Disappointingly, only 29% discussed the limitation of the proposed method, a finding that corroborates our claim that evaluations are too focused on reporting superior performance rather than fully characterizing system outputs. As a positive example, Kim et al. [2021] report negative results on out-of-distribution performance, encouraging future researchers to work on making their proposed method more robust.

On a positive note, a majority of papers (57%) report metrics from different categories instead of only relying on lexical overlap. In most such cases, the categories were metrics that measure similarity to a reference and diversity among outputs. However, some also developed metrics to specifically measure what is being claimed. For example, Lyu et al. [2021] work on lexical consistency for document-level MT, which they derive a metric from and use alongside other metrics to validate their specific claims. About 20% of papers provide additional breakdowns of the results, report on non-i.i.d. test sets, conduct error analyses, or demonstrate a causal effect of input features. These are especially helpful when the analysis is motivated by problem-specific needs. For example, Krishna et al. [2021] investigate the generation of doctors' notes from conversations and analyze the performance in the presence of simulated speech recognition errors.

While 29% of papers point out issues in the datasets they use or introduce, we found only one paper that contributed to the data documentation, leaving future researchers to rediscover the same issue(s). Moreover, only 3/13 papers that point out issues actually work toward solving them and release updates to the dataset. As discussed above, this is an area where normalizing contributing documentation and releasing updates would have beneficial effects for future work with these datasets.

Of the papers that report human evaluation results (73%), 82% state *what* is being measured, although the documentation of *who* is evaluating is still lacking (58%). We did not find a single paper that estimated how many annotations should be collected, and most opted for the "typical" 100 data points which, as pointed out above, may be insufficient [van der Lee et al., 2021]. Similarly, only 25% and 39% of papers assess the annotations and/or the annotators and only 21% discuss what background knowledge was required to participate in an evaluation.

The aspect that is lacking the most is the release of data. Though many papers released datasets or code to reproduce their models, almost none released model outputs or their human evaluation data. This can lead to issues when new papers are unable to compare using the same metrics environment, something that 37% of papers did not do. Moreover, it can significantly slow evaluation research due to a lack of data to annotate or human annotations to compare to.

Overall, our analysis demonstrates that there is much room for improvement in NLG evaluation, but it also shows that we are not starting at zero. While none of the papers reached 100%, which may be an overly ambitious goal, many reached 40% or higher, meaning that they already included many of our suggestions. We hope these best practices serve as a useful resource for researchers when designing and documenting NLG evaluations and for reviewers when evaluating NLG work.

References

Anja Belz and Albert Gatt. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P08-2050.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.inlg-1.24.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://www.aclweb.org/anthology/Q18-1041.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://www.aclweb.org/anthology/2020.acl-main.463.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385.

Daniel Deutsch and Dan Roth. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.conll-1.24.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Trans. Assoc. Comput. Linguistics*, 9: 774–789, 2021. URL https://transacl.org/ojs/index.php/tacl/article/view/2713.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan, October–November 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-8652. URL https://www.aclweb.org/anthology/W19-8652.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 04 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00373. URL https://doi.org/10.1162/tacl_a_00373.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. URL https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/fulltext.

Sebastian Gehrmann. *Human-AI Collaboration for Natural Language Generation with Interpretable Neural Networks*. PhD thesis, Harvard University, 2020. URL https://dash.harvard.edu/bitstream/handle/1/37365160/GEHRMANN-DISSERTATION-2020.pdf?sequence=1.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, and et al. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.10. URL https://aclanthology.org/2021.gem-1.10.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text, 2022. URL https://arxiv.org/abs/2202.06935.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-demos.6. URL https://aclanthology.org/2021.naacl-demos.6.

Yvette Graham. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1013. URL https://www.aclweb.org/anthology/D15-1013.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.413. URL https://aclanthology.org/2021.findings-acl.413.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.inlg-1.23.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://www.aclweb.org/anthology/2020.acl-main.560.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. GENIE: A leaderboard for human-in-the-loop evaluation of text generation. *CoRR*, abs/2101.06561, 2021. URL https://arxiv.org/abs/2101.06561.

Jihyuk Kim, Myeongho Jeong, Seungtaek Choi, and Seung-won Hwang. Structure-augmented keyphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2657–2667, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.209. URL https://aclanthology.org/2021.emnlp-main.209.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.384. URL https://aclanthology.org/2021.acl-long.384.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.360. URL https://www.aclweb.org/anthology/2020.findings-emnlp.360.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=mPducS1MsEK.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.262. URL `https://aclanthology.org/2021.emnlp-main.262`.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL `https://www.aclweb.org/anthology/2020.acl-main.173`.

Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.11. URL `https://aclanthology.org/2021.gem-1.11`.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In danah boyd and Jamie H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229. ACM, 2019. doi: 10.1145/3287560.3287596. URL `https://doi.org/10.1145/3287560.3287596`.

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P10-1056`.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL `https://www.aclweb.org/anthology/W18-6319`.

Anna Rogers and Isabelle Augenstein. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.112. URL `https://www.aclweb.org/anthology/2020.findings-emnlp.112`.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.647. URL `https://www.aclweb.org/anthology/2020.emnlp-main.647`.

Anastasia Shimorina and Anya Belz. The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in NLP. *CoRR*, abs/2103.09710, 2021. URL `https://arxiv.org/abs/2103.09710`.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.156. URL `https://aclanthology.org/2021.eacl-main.156`.

Amanda Stent, Matthew Marge, and Mohit Singhai. Evaluating evaluation methods for generation in the presence of variation. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005, Proceedings*, volume 3406 of *Lecture Notes in Computer Science*, pages

341–351. Springer, 2005. doi: 10.1007/978-3-540-30586-6_38. URL https://doi.org/10.1007/978-3-540-30586-6_38.

Craig Thomson and Ehud Reiter. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.inlg-1.22.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan, October–November 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-8643. URL https://www.aclweb.org/anthology/W19-8643.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151, 2021. doi: 10.1016/j.csl.2020.101151. URL https://doi.org/10.1016/j.csl.2020.101151.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.inlg-1.14.

A Annotation Instructions

Make informed evaluation choices and document them

- Evaluate on multiple datasets: Select yes if the paper reports results on more than one dataset. Select N/A if the paper explicitly states that there is only one dataset available for the addressed task.
- Motivate dataset choice(s): Select yes if the paper states why each particular dataset was chosen. If the only reasoning is that previous work uses it, select no. If the paper introduces a dataset, select N/A.
- Motivate metric choice(s): Select yes if the paper states why each particular metric was chosen. If the only reasoning is that previous work uses it, select no.
- Evaluate on non-English language: If at least one of the evaluated datasets includes non-English language, select yes.

Measure specific generation effects

- Use a combination of metrics from at least two different categories: Select yes, if the automatic evaluation results include at least two metrics from different families (e.g., one QA-based one and one lexical one). Reporting ROUGE and BLEU would not count while ROUGE and BLEURT would.
- Avoid claims about overall "quality": Select no if **the abstract** of the paper reports improvements generally and not in terms of specific generation aspects (e.g., "we outperform baselines")
- Discuss limitations of using the proposed method: Select yes, if there is at least one paragraph dedicated to the limitations of the proposed method in the results or discussion section or as its own section.

Analyze and address issues in the used dataset(s)

- Discuss or identify issues with the data: Select yes, if there is at least a mention of problematic artefacts with the data or what or who it represents.

- Contribute to the data documentation or create it if it does not yet exist: Select yes, if the paper is accompanied by a data card or if there is a mention that original documentation was updated.
- Address these issues and release an updated version: Select yes, if the paper is accompanied by a release of updated data or points to a loader that retrieves the updated dataset. If the paper introduces a dataset, select N/A.
- Create targeted evaluation suite(s): Select yes, if the paper describes the creation of a fine-grained breakdown of subpopulations **or** multiple training or test splits.
- Release evaluation suite or analysis script: Select yes, if the resources in the previous points were released in the form of data or code.

Evaluate in a comparable setting

- Re-train or -implement most appropriate baselines: Select yes, if the paper explicitly mentions that it trains or implements baselines from prior papers.
- Re-compute evaluation metrics in a consistent framework: Select yes, if **all** the reported scores were computed by the authors or by another centralized framework (e.g., through upload to a leaderboard). If only a subset was recomputed, select no.

Select N/A for both questions above if a new dataset was introduced and the only one evaluated in the paper.

Run a well-documented human evaluation

- Run a human evaluation to measure important quality aspects: Select yes, if a human evaluation of any kind was conducted.
- Document the study setup (questions, measurement instruments, etc.): Select yes, if, at the minimum, the specific questions and the way that participants answer them are reported.
- Document who is participating in the study: Select yes, if, at the minimum, the annotation platform used and the number of participants are stated.

Produce robust human evaluation results

- Estimate the effect size and conduct a power analysis: Select yes, if any effect size estimate or power analysis is mentioned (we assume that not mentioning it implies it absence).
- Run significance test(s) on the results: Select yes, if the human annotation results are accompanied by a statistical significance test.
- Conduct an analysis of result validity (agreement, comparison to gold ratings): Select yes, if there is any kind of analysis of the quality of the human annotations themselves.
- Discuss the required rater qualification and background: Select yes, if the required knowledge of raters is discussed and compared to the qualifications selected for in the study.

Document results in model cards

- Report disaggregated results for subpopulations: Select yes, if the paper reports fine-grained results on subsets of the test set(s) (note that the paper does not need to introduce these breakdowns as in the point above).
- Evaluate on non-i.i.d. test set(s): Select yes, if there is an evaluation on a non-i.i.d. test set. If the paper does not specifically mention this fact, select no (i.e., if the used dataset has such a test set but this is not mentioned).
- Analyze the causal effect of modeling choices on outputs with specific properties: Select yes, if the results include a breakdown that allow for insights of the form "if input has feature X, model output has Y". An ablation study counts as a yes, **if** the ablation focuses on feature representations (i.e. what data a model sees), but not if the ablation is on model architecture choices.
- Conduct an error analysis and/or demonstrate failures of a model: Select yes, if there is any kind of error analysis or qualitative samples of where the model fails.

Release model outputs and annotations

In this section, select yes, if the paper is accompanied by data releases that include the following.

- Release outputs on the validation set
- Release outputs on the test set
- Release outputs for non-English dataset(s): Select N/A if the paper does not include evaluation on any non-English data.
- Release human evaluation annotations

B Limitations

There are a few limitation of this analysis setup. (1) Due to the phrasing as recall-oriented prompts, nuanced errors pointed out in earlier sections are implicitly ignored. For example, “Document the study setup” is marked as positive even if the exact definition of each measurement category is not provided. The lack of providing a definition was identified as a source of confusion by Howcroft et al. [2020]. In other cases, our prompts may not be covering all possibilities. For example, a study that releases not an improved version of a corpus, but instead a tailored pretraining set would not count as “Address dataset issues and release an updated version”. (2) Each paper is only annotated by one co-author of this survey (after ensuring that the annotating author does not have a conflict of interest). This means that there could be misunderstandings of the different dimensions. We tried to address this problem by refining definitions when unclear points arose and by discussing the definitions before starting the annotation which led to the instructions above. Nevertheless, the exact percentage results may differ from the ground-truth by a few points and we thus consider only the overall trends when interpreting the results. (3) We are not releasing our annotations. To protect the identity of authors of papers with flawed evaluation processes according to our analysis, we will not release the data which may hinder reproducibility. We highlight a few positive examples in Section 4.

Implementing and popularizing these changes in the community will require several changes to peer review processes. First, we should encourage authors to submit resource papers. As Rogers and Augenstein [2020] point out, resource papers are already underappreciated and increasing what counts as acceptable documentation for a resource paper may lead to fewer such papers being written. Second, authors and reviewers need to move from claiming empirical improvements toward a more rigorous documentation of how those were achieved. Modeling papers often include deliberations why certain architecture choices were made, but the choice of which datasets to evaluate on or which metrics are being used rarely move beyond “other people use it”. By the same logic, reviewers may be hesitant to accept claims when a model is not evaluated on the standard flawed datasets. As discussed in this work, many of the standard practices should be reconsidered and we thus need more elaboration on these choices. Third, we encourage researchers to focus on specific phenomena, rather than overall quality. Instead of treating NLG models or metrics as “one big problem”, we encourage work on more specific aspects, say, logical consistency in dialog, or aggregations in table-to-text generation. We further encourage researchers to use task-specific metrics and be upfront with the trade-offs, and we encourage reviewers to expect and accept more nuanced claims and contributions while discouraging claims about the overall quality of a system. Finally, to support this research, we should encourage re-training and/or re-implementing prior work for the most appropriate benchmark task(s) and evaluation process when necessary.