Human Evaluation Reproduction Report for *Data-to-text Generation with Macro Planning*

Mohammad Arvan and **Natalie Parde**
University of Illinois Chicago, USA
{marvan3,parde}@uic.edu

Abstract

This paper presents a partial reproduction study of *Data-to-text Generation with Macro Planning* by Puduppully and Lapata (2021). This work was conducted as part of the ReproHum project, a multi-lab effort to reproduce the results of NLP papers incorporating human evaluations. We follow the same instructions provided by the authors and the ReproHum team to the best of our abilities. We collect preference ratings for the following evaluation criteria in order: conciseness, coherence, and grammaticality. Our results are highly correlated with the original experiment. Nonetheless, the presented results may be insufficient to conclude that the system proposed and developed by the original paper is superior compared to other systems. We suspect that combining our results with the three other reproductions of this paper through the ReproHum project will paint a clearer picture. Overall, we hope that our work is a step towards a more transparent and reproducible research landscape.

1 Introduction

Recent efforts have advanced the quality of automatic evaluation metrics, but these metrics still suffer from many shortcomings and flaws (e.g., a lack of correlation between scores and human judgments, such as that reported by Belz and Reiter (2006), Reiter and Belz (2009), Schluter (2017), Novikova et al. (2017), Post (2018), and van der Lee et al. (2019), among others) that render reliance on them less than ideal. Human evaluation eliminates most of these concerns, making it central to evaluating many machine learning, and in particular natural language processing (NLP), approaches. Nevertheless, evaluating the quality of algorithms and models using human raters still raises several unique challenges that can discourage researchers from doing so. For example, one prohibiting factor is cost: while automated metrics can be used repeatedly, essentially free of charge, human evaluations require the recruitment of paid raters with appropriate background knowledge or skillsets. The costs associated with this often force researchers only to evaluate a limited number of samples when conducting human evaluations, using crowd-sourcing platforms such as Amazon Mechanical Turk (AMT).[1] The use of crowd-sourcing platforms as a primary vehicle for subject recruitment can raise its own issues, as has been extensively documented by others even outside of the NLP research community (Goodman et al., 2013; Zhou and Fishbach, 2016; Arditte et al., 2016).

There have been many efforts to understand and mitigate the risks associated with human evaluation. Common practices include measuring inter-annotator agreement, calculating the power laws to select an appropriate sample size, and using statistical tests to measure the significance of the results (Wiebe et al., 1999; Snow et al., 2008; Pustejovsky and Stubbs, 2012; Dror et al., 2018; van der Lee et al., 2019). Undoubtedly, these practices further boost confidence in the results of human evaluation. However, they focus on pre-and post-analysis without providing insight into the human evaluation process. The lack of a systematic process for human evaluation has become a major concern in the last few years (Shimorina and Belz, 2021). Therefore, one may suggest that efforts to document and evaluate the human evaluation process are the next logical step to further improve the quality of human evaluation results without introducing any additional cost. This increased transparency and scrutiny is aligned with the goals of open science, will improve reproducibility, and will help the community to conduct higher-quality research.

From a broader perspective, concerns regarding scientific reproducibility are not new. In fact,

[1] https://www.mturk.com

the term *reproducibility crisis* has been used to describe the widespread barriers and inattention to reproducibility in many scientific fields (Baker, 2016; Wieling et al., 2018; Pineau et al., 2019; Belz et al., 2021; Pineau et al., 2021). With the increasing prominence of supervised machine learning methods that rely on empirical evidence in contemporary research, the importance of having reproducible results has become more important than ever. A global movement to promote increased reproducibility standards is gaining momentum (UNESCO, 2021), with the United Nations Educational, Scientific and Cultural Organization (UNESCO) taking a prominent role by underlining the value of open science with increased scrutiny and reproducibility as one of its main pillars. Ultimately, we can address reproducibility concerns by actively and systematically analyzing the current state of affairs, finding flaws, and proposing solutions (Belz et al., 2020; Sinha et al., 2021; Nature, 2022; Belz et al., 2022; ACL, 2022; Deutsch et al., 2022).

Over the last few years, many researchers have attempted to address the reproducibility crisis in NLP, often through meta-analyses and reproducibility studies of papers using automated metrics (Olorisade et al., 2017; Raff, 2019; Arvan et al., 2022a,b). Much less attention has been given to reproducibility studies of papers using human evaluations, mainly due to the additional complications of doing so (Belz et al., 2023). The ReproHum project aims to address this by conducting a large-scale, multi-lab reproducibility study of 50+ NLP papers incorporating human evaluations. As a participating lab in the ReproHum project, we were assigned a human evaluation experiment from *Data-to-text Generation with Macro Planning* by Puduppully and Lapata (2021). In this paper, we present our attempt to reproduce the results of that experiment. Thanks to the efforts of Puduppully and Lapata (2021) and the organizers of ReproHum, we were able to access most of the information required to reproduce our assigned experiment.

2 Background

Reproduction approaches were standardized across the ReproHum project, as summarized in this section (§2.1). We also present relevant evaluation details from the paper itself (§2.2), and we provide additional information from the paper's authors that was not included in the original paper itself but was necessary for reproducing the results (§2.3).

2.1 Common Approach to Reproductions

As a participating lab in the ReproHum project, we were provided with the following materials: (a) a document containing a common approach to reproduction, (b) the paper and the data required to reproduce the given experiment, and (c) a document containing all other additional information. We did not communicate with the authors directly. Instead, all communication was done through the ReproHum organizers. This decision was made to ensure consistency across reproductions and prevent authors from inadvertently influencing the reproduction process. It also enabled complete documentation of the process.

The document providing the common approach to reproductions offered a general overview of the process of reproducing a human evaluation experiment. The document was divided into two sections: one containing information for processes prior to the reproduction, and the other containing information for processes during and after the reproduction. The first section instructed us to familiarize ourselves with the paper and the experiment, and to calculate the amount of compensation required for crowd workers.[2] We were also asked to follow our own institutional guidelines regarding conducting human evaluation experiments. In our case, this involved applying for Institutional Review Board (IRB) approval at our own university (the University of Illinois Chicago). All outcomes of our reproduction were then achieved adhering to our approved IRB protocol.

The second section of the common approach focused on the reproduction process itself and subsequent data analyses. We were asked to fill out a Human Evaluation Data Sheet (HEDS) for each task. The HEDS is a spreadsheet that contains information about the task, the crowd workers, and the collected responses. Using this spreadsheet, we identified error types and created a side-by-side presentation of the results, findings, and conclusions to further assess the degree to which the reproduced outcomes matched the paper's original findings.

2.2 Evaluation Details from the Paper

Paper Summary. In our assigned paper, *Data-to-text Generation with Macro Planning*, Puduppully and Lapata (2021) augment a neural model

[2]Crowd workers providing annotations for ReproHum reproductions were all recruited from AMT using a single, centralized account.

with a macro planning stage for the task of data-to-text generation. This task aims to generate natural language that describes input data such as tabular data (e.g., databases of records or accounting spreadsheets) or structured data (e.g., knowledge graphs or semantic networks). The performance of end-to-end neural models has effectively rendered older techniques obsolete, but more modern models are far from perfect. The authors report that major issues including imprecision, hallucination, and poor context selection and document structuring plague modern models for this task. To address these issues, the authors propose the usage of macro planning, the high-level organization of information and how it should be presented. The authors highlight the current limitation of existing datasets for data-to-text generation using this approach, but note that nonetheless the expected output of these datasets is structured into several paragraphs, which can be used to define paragraph plans. Methodologically, the authors present a two-step pipeline for implementing their approach: first, a macro plan is generated using the training data, and then the plan is fed to a text generation model.

The authors use the RotoWire (Wiseman et al., 2017) and MLB (Puduppully et al., 2019) datasets to train and evaluate their proposed approach. Both datasets contain structured data about basketball and baseball games, respectively, with information pertaining to game statistics and summaries. They conducted human evaluation alongside automatic evaluation and empirically demonstrated that their generated text was more factual, coherent, and fluent compared to existing state-of-the-art models. Although their evaluation consists of both automatic evaluation and human evaluation, our focus is on the human evaluation part of their work. The human evaluation was performed through a comparative study of gold-standard output and four other systems, including theirs. Besides the model proposed by the authors (Macro), the other systems were: 1) a template-based generator (Templ), 2) ED+CC, which was the best performing system from an earlier study (Wiseman et al., 2017), and 3) the state-of-the-art model (RBF-2020) at the time of publication of the original paper (Rebuffel et al., 2020).

Human Evaluation. To conduct their human evaluation, Puduppully and Lapata (2021) used AMT. To ensure the acceptable quality of received responses, the authors required that workers had at least a 98% approval rate across at least 1000 previously completed tasks. Furthermore, they limited the locations of crowd workers to English-speaking countries (US, UK, Canada, Ireland, Australia, and New Zealand). The human evaluation was split into two tasks, with the first focusing on the number of supporting and contradicting facts in the game summaries and the second evaluating the quality of the generated text based on coherence, grammar, and conciseness. Our main objective was to reproduce the second task.

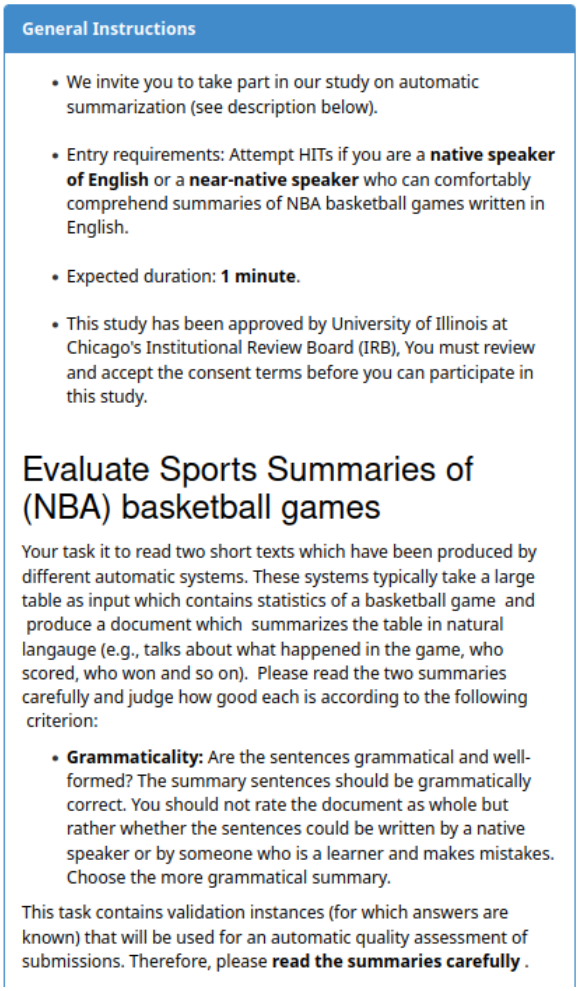

Figure 1: Instructions given to AMT workers for this task.

The second task elicited workers' preferences by asking them to compare two randomly selected summaries. Figures 1 and 2 illustrate the instructions and the input regions that the crowd workers used to respond. We used exact replicas of these in our reproduction (described later). The authors used Best-Worst Scaling (Louviere and Woodworth, 1991; Louviere et al., 2015) to present the results. The score for each system was calculated

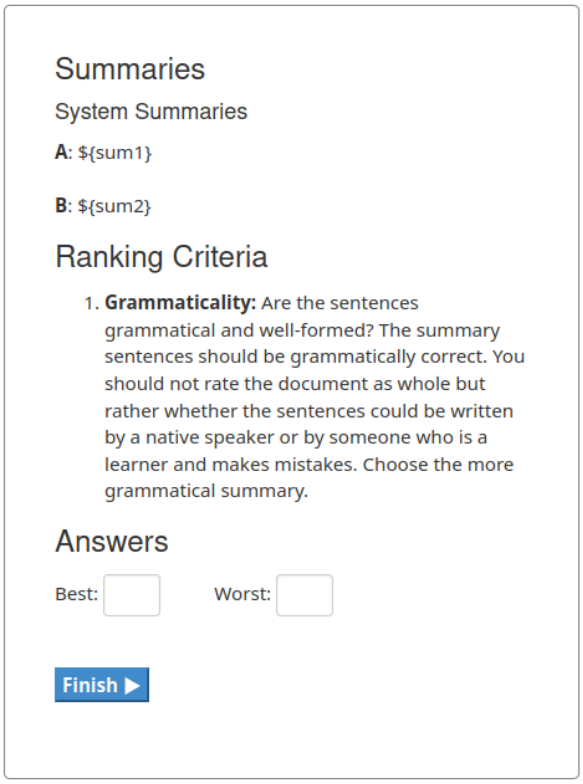

Figure 2: Specific input regions that AMT workers used to rank criteria associated with system summaries.

by subtracting the number of times the system was selected as the worst from the number of times it was selected as the best, divided by the total number of appearances of the system. The output of the four competing systems and gold output were divided into ten pairs of summaries. The evaluation criteria were grammar, coherence, and conciseness. Each pair was presented to three crowd workers to collect three distinct preference ratings per pair. Overall, the authors evaluated the system on the basis of 40 summaries (20 per dataset) and ten system pairs. With three evaluation criteria and three raters for each task, this meant that 3,600 preference ratings were solicited overall. The authors reported that 206 crowd workers overall participated in this task.

2.3 Additional Evalution Details from the Authors

Although we did not communicate with the authors directly, we were provided with a document containing additional information about the human evaluation process to support our reproduction. This information was acquired through correspondence between the authors and the ReproHum project team. The document contained information about the task setup, the instructions provided to the crowd workers, and the quality control measures that were employed. The ReproHum organizers mediated these correspondences to prevent undue influences to the reproduction process and to ensure that any communication between the authors and the reproduction team was documented. An additional practical motivation for this was that, as previously mentioned, two teams were assigned to reproduce each experiment—in requiring individual teams to refer to this document rather than correspond with the authors directly, the ReproHum organizers sought to maintain a level of consistency between the two teams.

The original authors were exceptional in providing additional information required to reproduce the experiments. For example, they granted us access to the original forms used in AMT to collect the responses. They also noted that while each task was assigned to three distinct crowd workers, the crowd workers had the option to accept multiple tasks. The authors also mentioned an exclusion criterion for the crowd workers to ensure the quality of the collected responses.

3 Methods

Our methods for reproducing the paper were as follows. We followed the same instructions provided by the ReproHum team to the best of our abilities, even following the exact same order of evaluation criteria as the other team. Specifically, we collected preference ratings for our evaluation criteria in the following order:

1. Conciseness
2. Coherence
3. Grammar

Each criterion was split into four mini-batches, each of which contained a quarter of the total number of tasks. The original authors incorporated attention checks to ensure the quality of received responses, by defining a set of conditions that (if met) would signal that the crowd worker should be excluded from the rest of the tasks. These exclusionary conditions were limited to the first two criteria (conciseness and coherence). For conciseness, they annotated and excluded the comparisons between all pairs except those involving the output generated by the template-based system. Since they

Model	Original			Ours		
	Gram	*Coher*	*Concis*	*Gram*	*Coher*	*Concis*
Gold	38.33	46.25*	30.83	14.17	12.50	5.83
Templ	-61.67*	-52.92*	-36.67*	-23.33*	-20.00*	-5.83
ED+CC	5.0	-8.33	-4.58	-8.33	-7.50	-5.00
RBF-2020	13.33	4.58	3.75	9.17	9.17	0.83
Macro	5.0	10.42	6.67	8.33	5.83	4.17

Table 1: Comparison of ROTOWIRE performance metrics. *Gram*, *Coher*, and *Concis* correspond to grammar, coherence, and conciseness, respectively. * indicates a statistically significant difference ($p < 0.05$) between Macro and the other systems. Note that the **Original** column numbers are from Table 5 of the original paper, while the **Ours** column numbers are from our reproduction.

no longer had access to the annotated exclusion criteria, we had to slightly diverge from the original process. As an alternative, we followed the instructions provided by the ReproHum team and limited the exclusion to pairs involving the gold output and one of the systems other than the template-based system. Specifically, the ReproHum team utilized NLTK[3] to compute an n-gram-based similarity score. The difference between the gold score and the system score was used to select 12 pairs with the highest difference. If any of the crowd workers rated one of these very different system outputs as superior to gold output, they were excluded from the rest of the tasks.

The exclusion process based on ratings of coherence was simpler than that used for ratings of conciseness. For coherence, if a crowd worker selected the template system output as superior to the gold output they were excluded from the rest of the tasks. Since we conducted our experiment after the other team assigned to this paper had finished their reproduction, workers excluded from the first team's study were also excluded from ours. Workers were paid for all tasks that they completed regardless of whether they were excluded. We paid workers \$0.22 per task, compared to \$0.15 in the original paper. This difference was due to adjustments for inflation and local minimum wage.

4 Results

Our results are summarized in Table 1. The results were computed using 1800 responses collected through twelve mini-batches (four for each of the three evaluation criteria). Each batch took approximately a day to finish collecting all responses. Overall, 262 crowd workers participated in this task.

[3] https://www.nltk.org

While the original study reported Krippendorff's $\alpha = 0.47$, ours was much worse ($\alpha = -0.011$). Note that the original authors calculated this coefficient using the results on both datasets; however, we computed our results using half the number of responses they used. The feedback we received from the crowd workers was positive.

We can observe from the results that the magnitude of difference reported between conditions in the original study's results is much higher than ours. For example, when evaluating grammaticality, the original study reports a best-worst scaling (BWS) score of -61.67 for the template system (the lowest score reported among all conditions), while ours is -23.33 (the lowest score reported among all conditions in our reproduction). Similarly, for coherence, our BWS score of 12.50 is much smaller than the reported BWS=46.25. We utilized the same statistical significance test as the original study (a one-way ANOVA with post-hoc Tukey HSD tests). The results of this test suggest that only two conditions (the Template system's scores for grammar and coherence) yield results with statistically significant differences from the Macro system. This is a different finding from the original study, which reported statistically significant different results for four measures. These measures were Templ for grammar, coherence, and conciseness, and Gold for coherence.

In our analyses of the observed errors, we found a high level of similarity between the original experiment and our reproduction. We used Pearson's r and Spearman's ρ to measure the correlation between the two experiments. With Pearson's $r = 0.90$ and Spearman's $\rho = 0.83$, we can conclude that the outcomes from the two experiments are highly correlated. In other words, in spite of

the differences explained and observed between the two studies, our results do not invalidate the original study's findings.

5 Discussion

To discuss the implications of our findings, we first reiterate the contributions of the original study and the scope of our reproduction. Puduppully and Lapata (2021) presented a novel technique with the goal of improving the quality of data-to-text generation. They used a combination of automatic and human evaluation methods to show that their approach was superior to existing state-of-the-art models on two datasets, RotoWire and MLB. The scope of our reproduction was limited to the second human evaluation task reported in their paper, examining the quality of generated text based on coherence, grammaticality, and conciseness. Furthermore, we only reproduced the results on the RotoWire dataset. To provide a better perspective, MLB dataset, is larger (nearly ten times as many tokens) than the RotoWire dataset. Hence, we cannot form conclusive judgments based on a full reproduction of this experiment; rather, we focus on a subset of it.

Thus, our outcomes are currently inconclusive but promising, with evidence of a high level of similarity between our findings and those originally reported. Through our focus on the results that are available, we do not believe that there is enough evidence to claim that the Macro system proposed and developed by the original paper is superior compared to other systems. However, we believe that combining our results with the three other reproductions of this paper through the ReproHum project will paint a clearer picture. Therefore, we leave the final judgment to the ReproHum team.

Regarding the reproduction process itself, we found that many details required to successfully reproduce the original work were missing from the paper. We believe that this is likely due to many factors associated with the current NLP research climate, including an overemphasis on novelty, formatting, and paper length, that are all beyond the original authors' control. Thanks to the cooperation of the authors, we were able to find answers to the most important questions. We underscore that this level of communication is hard to find. Unfortunately, there are still little to no guidelines regarding the long-term support of research artifacts and files once studies have been published. It is hard to imagine the contemporary machine learning and natural language processing research landscapes without empirical studies driving them forward. At the same time, perhaps conferences and journals should consider potential avenues for collecting technical details beyond what has been made available in the paper itself. Another option is to further encourage the publication of reproduction studies in primary publication venues.

6 Conclusion

In this work, we have presented our attempt to reproduce the human evaluation of one experiment from *Data-to-text Generation with Macro Planning* by Puduppully and Lapata (2021). Overall, with Pearson's $r = 0.90$ and Spearman's $\rho = 0.83$ when comparing outcomes of the original study and our reproduction, we can conclude that when reproducing the experiment as described in the paper we observe highly correlated results. Nonetheless, we believe that without the help and cooperation of the original authors, we might have observed a different outcome. We note that the reproduced results in this work are only a portion of the results presented in the original paper. Therefore, concluding that the claims made by the original study are valid at this point would be premature. We leave the final judgment to the ReproHum team.

Acknowledgments

We are immensely grateful to Ratish Puduppully and Mirella Lapata, the authors of the original paper, for their invaluable work and exceptional responsiveness in providing additional information and support throughout this reproduction project. Their cooperation and guidance have been instrumental in ensuring the accuracy and fidelity of our work. We also extend our appreciation to the ReproHum project team, including Anya Belz, Ehud Reiter, Craig Thomson, and Maja Popović, for their collaboration and expertise, which have enriched this endeavor. Furthermore, we would like to express our sincere thanks to the Engineering and Physical Sciences Research Council (EPSRC) for their generous grant support (EP/V05645X/1), without which this project would not have been possible. The collective efforts of all involved have been crucial in shaping the success of this reproduction, and we are truly thankful for their support and contributions.

References

ACL. 2022. ACL Responsible NLP Research.

Kimberly A Arditte, Demet Çek, Ashley M Shaw, and Kiara R Timpano. 2016. The importance of assessing clinical phenomena in mechanical turk research. *Psychological assessment*, 28(6):684.

Mohammad Arvan, Luís Pina, and Natalie Parde. 2022a. Reproducibility in computational linguistics: Is source code enough? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2350–2361. Association for Computational Linguistics.

Mohammad Arvan, Luís Pina, and Natalie Parde. 2022b. Reproducibility of exploring neural text simplification models: A review. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 62–70, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604).

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2020. ReproGen: Proposal for a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 232–236. Association for Computational Linguistics.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 381–393. Association for Computational Linguistics.

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Huiyuan Lai, Chris van der Lee, Emiel van Miltenburg, Yiru Li, Saad Mahamood, Margot Mieskes, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Pablo Mosteiro Romero, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Daniel Deutsch, Yash Kumar Lal, Annie Louis, Pete Walsh, Jesse Dodge, and Niranjan Balasubramanian. 2022. 2022 North American Chapter of the Association for Computational Linguistics Reproducibility Track.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1383–1392. Association for Computational Linguistics.

Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. 2013. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 355–368. Association for Computational Linguistics.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Nature. 2022. Nature's Reporting standards and availability of data, materials, code and protocols.

Jekaterina Novikova, Ondrej Dusek, 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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2241–2252. Association for Computational Linguistics.

Babatunde Kazeem Olorisade, Pearl Brereton, and Peter Andras. 2017. Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist. *J. Biomed. Informatics*, 73:1–13.

Joelle Pineau, Koustuv Sinha, Genevieve Fried, Rosemary Nan Ke, and Hugo Larochelle. 2019. Iclr reproducibility challenge 2019. *ReScience C*, 5(2):5.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research(a report from the neurips 2019 reproducibility program). *J. Mach. Learn. Res.*, 22:164:1–164:20.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2023–2035. Association for Computational Linguistics.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Trans. Assoc. Comput. Linguistics*, 9:510–527.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning - a Guide to Corpus-Building for Applications*. O'Reilly.

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5486–5496.

Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. A hierarchical model for data-to-text generation. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035 of *Lecture Notes in Computer Science*, pages 65–80. Springer.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Comput. Linguistics*, 35(4):529–558.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 41–45. Association for Computational Linguistics.

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

Koustuv Sinha, Jesse Dodge, Sasha Luccioni, Jessica Forde, Sharath Chandra Raparthy, François Mercier, Joelle Pineau, and Robert Stojnic. 2021. ML Reproducibility Challenge 2021.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 254–263. ACL.

UNESCO. 2021. UNESCO recommendation on open science.

Janyce Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999*, pages 246–253. ACL.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Comput. Linguistics*, 44(4).

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2253–2263. Association for Computational Linguistics.

Haotian Zhou and Ayelet Fishbach. 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of personality and social psychology*, 111(4):493.