ReproGen: Proposal for a Shared Task on Reproducibility of Human Evaluations in NLG

Anya Belz
University of Brighton, UK
a.s.belz@brighton.ac.uk

Shubham Agarwal
Heriot Watt University, UK
sa201@hw.ac.uk

Ehud Reiter
University of Aberdeen, UK
e.reiter@abdn.ac.uk

Anastasia Shimorina
Université de Lorraine / LORIA, France
anastasia.shimorina@loria.fr

Abstract

Across NLP, a growing body of work is looking at the issue of reproducibility. However, replicability of human evaluation experiments and reproducibility of their results is currently under-addressed, and this is of particular concern for NLG where human evaluations are the norm. This paper outlines our ideas for a shared task on reproducibility of human evaluations in NLG which aims (i) to shed light on the extent to which past NLG evaluations have been replicable and reproducible, and (ii) to draw conclusions regarding how evaluations can be designed and reported to increase replicability and reproducibility. If the task is run over several years, we hope to be able to document an overall increase in levels of replicability and reproducibility over time.

1 Introduction

Human evaluations play a central role in Natural Language Generation (NLG) (Reiter, 2018; Novikova et al., 2017), so it is of concern that we do not currently know to what extent their results are reproducible, hence whether they are reliable or not. Reproducibility is on many NLP researchers' minds at present. There have been workshops on replicability and reproducibility in NLP most years since 2015.[1] The Reproducibility Challenge has been running since 2018, initially in conjunction with ICLR'18 and ICLR'19 (Pineau et al., 2019), then at NeurIPS'19 (Sinha et al., 2020) and NeurIPS'20 (to appear). COLING'18 had a Reproduction Paper special category, for which it reported 35 submissions. NeurIPS'19 had a reproducibility programme comprising a code submission policy, a reproducibility challenge for machine learning (ML) results, and the ML Reproducibility checklist for submitted papers (Pineau et al., 2020) which has also been adopted by EMNLP'20 and AAAI'21. LREC'20 ran a reproducibility track (Branco et al., 2020). Other conferences have foregrounded reproducibility via calls, chairs' blogs, special themes and social media posts.

All this is against the wider background of what has been called a 'reproducibility crisis' (Baker, 2016) in science, where 70% of scientists report failing to reproduce someone else's results on at least one occasion, and over half report failing to reproduce own results. In NLP, 24.9% of attempts to reproduce own results, and 56.7% of attempts to reproduce another team's results, fail to reach the same conclusions (Mieskes et al., 2019).

Progress is being made in NLP regarding reproducibility, as can be seen from the long list of events and initiatives above. The habit of sharing code, data and supplementary material providing details about data, systems and training regimes is firmly established in the field, virtually all main events now encouraging and making space for it. Moreover, reproducibility is beginning to be addressed formally in the reviewing process, e.g. EMNLP'20 followed NeurIPS'19 in adding the ML Reproducibility Checklist (Pineau, 2020) to submission forms,[2] where authors had to indicate compliance with reproducibility criteria (although this was not used in selection decisions).

While progress is being made on many fronts, there is one big gap in efforts to achieve greater reproducibility, and that concerns human evaluation. If a paper complies with all of the NeurIPS'19/EMNLP'20 reproducibility criteria, it should be possible to reproduce metric results reported in it closely. However, any human evaluation results reported for the same system(s) in the same paper may or may not be reproducible, because the criteria say nothing at all about those.

[1] IJCAI'15 Workshop on Replicability and Reproducibility in NLP.

[2] https://2020.emnlp.org/call-for-papers

Proceedings of The 13th International Conference on Natural Language Generation, pages 232–236,
Dublin, Ireland, 15-18 December, 2020.

This shared task proposal is part of a wider effort to address reproducibility of human evaluations in NLG, a field in which they play a central role and which has always been wary of automatic evaluation metrics and their limitations (Reiter and Belz, 2009; Novikova et al., 2017; Reiter, 2018). Below we start by briefly diagnosing the reproducibility problem in NLG (Section 2), and looking at what the conditions are for reproducibility testing of results from human evaluation of NLG systems (Section 3). We then outline our ideas for a shared task that could help to understand and potentially address the problem (Section 4). We describe related research that has provided inspiration (Section 5), and conclude with next steps (Section 6).

2 Issues with Human Evaluation in NLG

In an era dominated by metrics and leaderboards, human evaluation can be an afterthought, often carried out and reported in a slapdash way with tiny numbers of evaluators and included to add a veneer of credibility. Data collected for a recent survey of 20 years of human evaluation in NLG (Howcroft et al., 2020) indicates that 33% of evaluations use fewer than 10 evaluators, and 22% use between 1 and 4, numbers so small that experiments are unlikely to yield meaningful results, or to be reproducible, in many contexts. The survey also revealed that the roughly 170 papers reviewed often provide woefully inadequate information about human evaluations: numbers of evaluators, experimental design, the quality criterion assessed, even system language, are often unclear. Researchers moreover use a wide variety of different quality criteria, with a startling 200-odd different quality criterion names found in the survey. Even when researchers do use the same criterion *name* they often don't use it with the same *meaning*, and vice versa (see also Van Der Lee et al. 2019).

Inter-rater and intra-rater agreement, two indicators of human evaluation reliability, are rarely reported. Amidei et al. (2019) surveyed 135 NLG papers from 2008-2018 and found that just 18% reported any annotator agreement, and where agreement scores were reported they were low, casting doubt on the reliability of results.

We're aware of one published attempt to reproduce someone else's human evaluation results in NLG: Cooper and Shardlow (2020), as part of REPROLANG, successfully reproduced system rankings in text simplification (however, with lower means). Another paper tested the ability of evaluators to reproduce their own evaluations in exactly the same experimental set-up which found that for some tasks and some evaluation instruments, evaluators struggled to reproduce their own evaluations (Belz and Kow, 2010). NLG has valued and trusted human evaluation perhaps more than any other NLP subfield, so it's concerning that we currently don't know if any given set of human evaluation results are reproducible, meaning we don't know whether we can, in fact, trust them.

3 Testing Human Evaluations for reproducibility

For results from human evaluations to be deemed reproducible, the first condition is that *the system outputs assessed in the human evaluation must be reproducible*, and the NeurIPS'19/EMNLP'20 criteria provide a pretty comprehensive set of conditions for that to be achievable. We take this as our starting point, i.e. we assume we have been successful in reproducing system outputs. In the shared task we will not try to reproduce system outputs, but start from existing sets of outputs (and inputs, where applicable), and try to reproduce the results of human evaluations performed on them.

3.1 Replicating experiments

The second condition is that *the experiment that produced the human evaluation results must be replicable* which means having access to detailed information about how the experiment was designed and run, but also that it is repeatable *in principle*. Belz et al. (2020) identify a set of 18 properties with associated value sets for characterising evaluations that are needed for replicability: in addition to defining quality criteria and evaluation mode, papers need to include, or give access to, full details of system outputs (number, how selected), evaluators (number, type, how selected), method for determining effect size and significance of findings, scale or other rating instrument (size, list or range of possible response values), how presented to evaluators, form of response elicitation, information given to evaluators, and experimental conditions. This level of detail is currently extremely rare in NLG papers (Van Der Lee et al., 2019; Howcroft et al., 2020).

However much detail is provided, it is bound to be an approximation rather than a complete specification. Aspects such as the lab environment, the

interface design and the exact training or instructions given are not normally reported, but may have considerable impact on results. More commonly described details may not be replicable either generally or for a particular team of researchers. For example, expert evaluators or proprietary software may not be accessible. An experiment may not be repeatable even in principle, e.g. if data protection laws or ethical regulations have changed, or if it was conducted in a one-off real-life context.

3.2 Reproducing results

The third condition is that *the replication must produce results that are the same as those produced by the original human evaluation, under the terms of a given frame of reference*. The latter needs to set out under what conditions, in terms of which aspects, two evaluations can be considered to have produced the same results. We know we can't demand that they be identical, given the limits to replicability discussed above, but we need some principled way of determining 'sameness'.

A related question is how to factor in variation in experimental design. Presumably one would want results to be very similar if the experiment is repeated in an identical manner in terms of the details listed in the previous section, with the same experimental software, and the same evaluators. But what about differences in user interface design? Or in the method for allocating test items to evaluators?

Variation in design, like similarity of results, is likely to be most usefully construed as a matter of degree, and reproducibility results reported in terms of (i) the level of variation in the experiment, and (ii) the level at which it has been possible to reproduce results. This would make it possible to characterise outcomes of different reproducibility tests in the same terms and make them comparable.

With the ReproGen Challenge we aim to start finding answers to the above questions, as a community of researchers, in a public process. Initially (Section 4), the task for participants will be to take available information about a human evaluation and try, with support from the original authors, to replicate experiments as closely as possible, then report the outcome. With multiple teams trying to reproduce results from the same papers, we will have a more complete view of the reproducibility of individual sets of results (in contrast to other studies which often make just one attempt, e.g. Open Science Collaboration 2015).

3.3 Interpreting reproducibility results

Following the above approach, the outcome of a reproduction attempt is not simply success or failure, it's a matter of how similar the two experiments were and how many aspects the attempt reproduced successfully. To facilitate these assessments, we will select, where possible, papers for reproduction that report system rankings, mean system level scores, significant differences, p-values, and effect sizes, and ask participants to report corresponding results. In addition, we will ask participants to document their reproduction attempts as precisely as possible including any gaps that had to be filled in.

Low reproducibility can have diverse causes: lack of detail about the original experiment, a flaw in the experimental design, or a problem with the evaluation task at the core of it (e.g. if participants find it too hard to score a given quality criterion).

Reproducibility outcomes from the proposed shared task will allow conclusions to be drawn about what kind of experiments are easier to reproduce, what the required level of information about experiments is to make them replicable, results from what types of experimental design are more, or less, reproducible, and how different aspects of experimental design and implementation affect reproducibility.

4 Organisation of Shared Task

We envisage ReproGen to have two tracks, one an 'unshared task' in which teams attempt to reproduce their own prior human evaluation results, the other a shared task in which teams try to reproduce the same prior human evaluation results. We envisage a fairly simple challenge first, where we nominate about five papers as replication targets for participants to choose from. Attendees can then either (A) replicate one of these experiments, or (B) replicate one of their own previous experiments:

A ***Main Reproducibility Track:*** For a shared set of selected human evaluations, participants attempt to reproduce their results, using published information plus extra detail provided by the authors (discussion with and support from the original authors e.g. played a big role in the Reproducibility @NeurIPS'19 challenge for ML results (Sinha et al., 2020)), and making common-sense assumptions where information is still incomplete.

B ***RYO Track:*** *R*eproduce *Y*our *O*wn previous

human evaluation results, and report what happened. Unshared task.

For the main track (A above), the plan is to ask authors to volunteer their papers for inclusion via an open call for expressions of interest. Authors will be invited to provide additional details and/or software to help teams reproduce the results. It's not clear how much this will help: while some studies indicate large increases from author help (Raff, 2019), other large-scale studies show no improvement at all (Klein et al., 2019).

Neither of the tracks would have winners or leaderboards in the normal shared-task sense. However, 'winners' in the main track would provide toplines of reproducibility for the included papers, and taken together, ReproGen Challenge contributions would help shed light on how to improve the reproducibility of human evaluations in NLG.

We expect teams to participate for a variety of reasons, ranging from researchers new to human evaluation (the Reproducibility Challenge @NeurIPS for ML results encouraged computer science courses to get their students to participate), to researchers experienced in human evaluation specifically interested in reproducibility.

Participation will have financial implications in the case of some papers. We will endeavour to keep such cost as low as possible by selecting mostly papers that were not expert or crowd-evaluated. We are also applying for funding to support crowd-based evaluations.

5 Related Work

Reproducibility investigations are commonly conducted in a closed project. For example, the Open Science Collaboration (2015) conducted 100 attempts to reproduce studies in psychology, mainly evaluating reproducibility using significance, p-values, effect sizes, and meta-analysis of effect sizes. They found substantial decreases from original to replication study for all three indicators, while just 39% of effects were subjectively rated to have reproduced the original result.

In contrast, the shared task framework ensures openness to all members of a research community. The Reproducibility Challenge @NeurIPS'19, focusing on ML results and metric scores, was organised as a 'live' challenge, where participants pick one of the NeurIPS accepted papers, and try to reproduce its ML results (Sinha et al., 2020). NeurIPS'19 authors were strongly encouraged to submit code and data, which 73% did, resulting in a 'codebase' the Reproducibility Challenge participants could choose from to participate in one of three tracks: (i) a baselines track (rigorous analysis of baseline results, re-implementing them if necessary), (ii) an ablation track (rigorous ablation experiments, modifying model and hyperparameters using the authors' code), and (iii) a replications track (replication of experiments in paper from scratch without using code from codebase).

The challenge was a big success, attracting 83 submissions after participants initially claimed 173 NeurIPS papers. The submissions were peer-reviewed as part of the NeurIPS reviewing process which relied heavily on the OpenReview platform, and 10 papers were selected for publication in ReScience C, an open access journal intended as a forum for replication work in computing science.

The one paper on reproducing human evaluation results in NLG mentioned above (Cooper and Shardlow, 2020) was part of the REPROLANG'20 initiative which followed on from two earlier, smaller-scale[3] LREC workshops on reproducibility and citation, and offered a shared task (Branco et al., 2020) which asked participants to reproduce results from one of 11 papers from different areas of NLP. While in the case of ten papers, the results up for reproduction were automatic scores, in one case they included human evaluation scores.[4]

6 Next Steps

With this shared task proposal we hope to engage the NLG community in a discussion about how best to design and organise the ReproGen Challenge. Following feedback and input, we will finalise the task specification and organisational aspects, expecting to be able to launch the task in 2021 for a pilot run with around five sets of human evaluation results up for reproduction.

We would hope that the ReproGen Challenge will both shed light on the reproducibility of current human evaluations in NLG, and allow conclusions about how evaluations can be designed and reported to increase reproducibility. Over repeated instances of the Shared Task, we hope to be able to document an overall increase in the reproducibility of new human evaluations in NLG.

[3] 4REAL2016 and 4REAL2018 had four papers each and one actual reproduction attempt.

[4] Task D.1: Text simplification: http://wordpress.let.vupr.nl/lrec-reproduction/

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.

Monya Baker. 2016. Reproducibility crisis. *Nature*, 533(26):353–66.

Anya Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*.

Anya Belz, Simon Mille, and David Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

David Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid Hasan, Saad Mahamood, Simon Mille, Sashank Santhanam, Emiel van Miltenburg, and Verena Rieser. 2020. 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*.

Richard A Klein, Corey L Cook, Charles R Ebersole, Christine Vitiello, Brian A Nosek, Christopher R Chartier, Cody D Christopherson, Samuel Clay, Brian Collisson, Jarret Crawford, et al. 2019. Many labs 4: Failure to replicate mortality salience effect with and without original author involvement.

Margot Mieskes, Karën Fort, Aurélie Névéol, Cyril Grouin, and Kevin Cohen. 2019. Community perspective on replicability in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 768–775, Varna, Bulgaria. INCOMA Ltd.

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.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251).

Joelle Pineau. 2020. The machine learning reproducibility checklist v2.0.

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 Fox, and Hugo Larochelle. 2020. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *CoRR abs/2003.12206*.

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems*, pages 5485–5495.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

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

Koustuv Sinha, Joelle Pineau, Jessica Forde, Rosemary Nan Ke, and Hugo Larochelle. 2020. NeurIPS 2019 Reproducibility Challenge. *ReScience C*, 6(2):#11.

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*, pages 355–368.