ReproHum #0124-03: Reproducing Human Evaluations of end-to-end approaches for Referring Expression Generation

Saad Mahamood
trivago N.V.
Düsseldorf, Germany
saad.mahamood@trivago.com

Abstract

In this paper we describe our attempt to reproduce a single human evaluation quality criterion of the human evaluation that was in conducted in the paper "NeuralREG: An end-to-end approach to referring expression generation". In particular, this paper describes the approach and challenges involved in reproducing the human evaluation as done by the original authors of the paper, the results obtained, and what insights we have gained from attempting this particular reproduction. Insights that we hope will enable refinements to both how human evaluations are documented by author(s) and enable better reproductions of NLP experiments in the future.

Keywords: human evaluation, NLP, neural REG, reproduction

1. Introduction

There has been significant interest in understanding the issues that prevent the reproduction and repeatability of human NLP evaluations. Efforts such as the ReproHum project[1] attempts to investigate human evaluations within NLP by systematically uncovering the extent of problems of reproducibility. Uncovering these issues is especially important within the field of NLP considering the significance of human evaluations, which are seen as the "gold standard" as compared to automatic metric based evaluations, which may not correlate well with human judgement (Belz and Reiter, 2006). Past research has indicated only a minority of systems can reproduce previously reported scores and systems due either to not working non-functional code or resource limits (Belz et al., 2021b). In fact some estimates place the percentage of papers being repeatable without any significant barriers as low as 5% and at 20% if the original author(s) help is sought (Belz et al., 2023). In addition to buggy code, other issues have been observed such as flaws within the user interface to collect evaluator responses, inappropriate exclusion of evaluators and/or data points, reporting flaws, and also ethical flaws (Thomson et al., 2024).

As part of the ReproHum multi-lab study (Belz and Thomson, 2024), multiple partner labs have come have come together to reproduce existing human evaluations experiments from a chosen set of human evaluations in published NLP research papers. Papers that were vetted by the organising committee to ensure that sufficient details in terms of materials (code, data, etc.) and evaluation procedures were present for a successful attempt at reproduction by a given partner lab. In addition to the original paper author(s) consent and co-operation was sought to enable the reproduction of human evaluations in their paper. Consecutively participating partner labs must follow a common reproduction approach to ensure consistency and comparability between different reproduction attempts.

This years reproduction experiment is a continuation of past years, which since 2021[2] has expanded the scope of reproduction experiments. Results from previous iterations have found the impact that different cohorts can have in the reproducibility of a given experiment (Belz et al., 2021a), or the need to lower cognitive loads for evaluators, which could potentially lead to be better reproducibility of results (Belz et al., 2022). In the 2023 edition there were three main challenges identified in trying to run reproduction results. The first was reproduction attempts encountering bugs, errors, and flaws, which were fixed differently by different reproducing authors. Secondly, reproducing authors chose different results to reproduce and report making comparability between results not possible. Finally, not all reproducing authors were able to adhere closely to the original experiment details with variations occurring such as using a different evaluation interface, or different number of evaluators (Belz and Thomson, 2023).

Based on the learnings from last year several changes have been implemented by the organis-

[1] ReproHum - https://reprohum.github.io

[2] ReproGen 2021 - https://reprogen.github.io/2021/
ReproGen 2022 - https://reprogen.github.io
ReproNLP 2023 - https://repronlp.github.io/2023

Proceedings of the HumEval @LREC-COLING 2024, pages 250–254
Turin, Italy. May 21st, 2024.

ers. There is now a revised and expanded common approach to reproduction that formalises that gives greater guidance on how the reproduction should be conducted and how the results should be reported to ensure greater comparability and standardisation between different reproduction attempts for the same paper.

In this paper we give a description of our attempt to reproduce human evaluations within the paper "NeuralREG: An end-to-end approach to referring expression generation" by Castro Ferreira et al. (2018) (section 2) and how the reproduction of the paper was conducted. We detail the challenges involved (section 3). We also detail the results obtained from the reproduction (section 4) and how they compare to the original results and the observations made by authors. Finally, we conclude with the learnings (section 5) that we have obtained based on the experiences of this reproduction experiment and describe improvements that would enable more robust reproductions of future NLP human evaluations.

2. Reproduction Experiment

In this reproduction experiment we were tasked with to reproduce human evaluations was "NeuralREG: An end-to-end approach to referring expression generation" (Castro Ferreira et al., 2018). The paper itself describes the creation of and an evaluation of an end-to-end neural approach for generating referring expressions, which then compared against two non-neural baseline models using the WebNLG dataset (Gardent et al., 2017). In particular, there are three neural variant systems that uses a different LSTM decoders tested by the authors and two non-neural variants:

- ***OnlyNames*** – A baseline non-neural model that leverages the similarity among the Wikipedia ID of an element and proper name reference to it. Basically, it replaces the underscores in a given Wikipedia ID for whitespaces.
- ***Ferreira*** – A second non-neural baseline model that leverages the Naive Bayes method to determine whether a given reference should be a proper name, pronoun, description, or demonstrative.
- **NeuralREG+**`Seq2Seq` – Leverages a decoding approach that models a given context vector for a given time step and concatenates the pre- and pos-context annotations averaged over time.
- **NeuralREG+**`CAtt` – A LSTM decoder that is augmented with an attention mechanism (Bahdanau et al., 2014), which for a given time step, used over the pre- and pos-context encodings.
- **NeuralREG+**`HierAtt` – Inspired by Libovický and Helcl (2017), this version implements a second attention mechanism in order to generate attention weights for pre- and pos-context summary vectors instead of concatenating them.

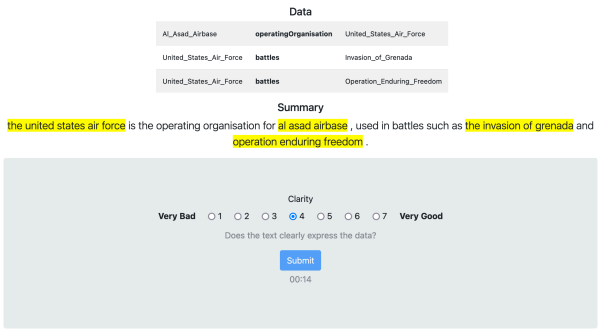

Figure 1: Evaluation interface used for rating the degree of clarity of a text containing generated referring expressions (highlighted in yellow).

Whilst these systems were evaluated using both automatic and human evaluations, the focus of this reproduction is solely on the human evaluation conducted by the original authors. In particular, the authors designed an intrinsic evaluation tasked that leveraged 24 randomly selected test WebNLG triplet instances and generated 6 target texts with referring expressions: The original (randomly selected) and five other referring expression version texts generated by each of the models described above. Using a latin square design, the authors created 144 different trials over 6 different list and designed the evaluation in a way that a given participant rated 24 trials, one for each of the 24 corpus instances, ensuring that participants saw an equal number of triplet set sizes and generated versions.

In the original experiment participants were asked to rate in a given trial three aspects for a given text containing referring expressions: *fluency*, *grammaticality*, and *clarity*. For the reproduction experiment we are tasked with only reproducing the *clarity* quality criteria aspect. Defined by the authors as whether the text clearly expresses the data. The quality criterion were rated by the participants using a seven point Likert scale. The task as done in the reproduction experiment is illustrated in figure 1, which shows a given set of triplets presented to the user in a tabular form and the text underneath with the generated referring expressions highlighted in yellow. Annotators are given 20 seconds to consider the data, the text with the generated referring expressions, and then give their ratings. This timer is unchanged from the original experiment even though the the number of quality criterion has been reduced from three to one.

Other changes to the user interface were limited in scope to accommodate ethical concerns or to update explanatory text to the fact that only one quality criteria aspect would evaluated instead of three. Changes included adding informed consent

Aspect	Original	Reproduction
Quality Criterion	*fluency, grammaticality, clarity*	*clarity*
Number of Items	144	144
Number of Systems	6	6
Number of Participants	60	60
Participants per Item	10	10
Items per Participant	24	24
Recruitment Platform	*Amazon MTurk*	*Prolific*
Compensation	*unknown*	*£12.00 per hour equivalent*
Participation controls	*unknown*	*none*
Age	*Average 36 years*	*Majority 18-24 years (43%)*
Gender Split	*27 females, 33 males*	*35 females, 25 males*
English Proficiency	*Native:* 44, *Fluent:* 14, *Basic:* 2	*Native:* 37, *Fluent:* 21, *Basic:* 2

Table 1: Methodological similarities & differences between the original and reproduction human evaluations.

information, amending the granularity of age information collected, and adding a more representative set of gender options.

3. Methodology & Challenges

Participants for the original experiment were recruited from Amazon Mechanical Turk, with 60 participants recruited and 10 assigned for each of the six lists. In the reproduction experiment, participants were recruited instead from Prolific[3] in agreement with the ReproHum organisers to ensure every reproduction experiment used a standardised crowd working platform. Whilst, the original experiment does not detail the degree of compensation given to participants, for the reproduction experiment participants were paid the equivalent of the UK living wage[4] of £12.00 per hour for their participation. Table 1 details the methodological and participatory similarities and differences between the two experiments. In terms of demographics, in the reproduction experiment the age is much younger that in the original experiment with 43% of participants reporting themselves to be between 18-24 years old and there is a greater proportion of participants identifying as female compared to the original experiment. For English proficiency, the distribution between the original and reproduction are fairly similar although with a slight more number of fluent instead of native English speakers.

The experimental data and user interface was taken from the original published source code repository[5]. The main evaluation interface, was a web application that was written in PHP with the purpose of handling collecting user responses and assigning users to an equal number of evaluation lists. However several challenges were encountered in attempting to reuse the original experimental data and user interface:

- The database structure was not available in the GitHub repository. As part of the reproduction this had to be recreated by interpreting the existing PHP code and through trial and error.
- The order of evaluations items was not defined for each list as this was was encoded in the not provided database.
- Lack of detailed version information for both the software used for the evaluation interface and the analysis code.

For the second point, whilst the start item for each of six lists was hard coded into the PHP code the order of subsequent items was not known. Therefore in coordination with the ReproHum organisers it was decided to randomise the order of items for each of the six lists. However, this change may have lead to a potential deviation from how the original experiment was conducted by the authors. Whilst, writing this report it was discovered that the code for the generate the trial lists was hidden in a python file that was used for computing the result statistics of the human evaluations.

In addition to setting up the reproduction experiment by using the original experiment's codebase a Human Evaluation Datasheet (HEDS) (Shimorina and Belz, 2022) was also completed[6]. The HEDS form records in a standardised way the properties of human evaluations to support comparability, meta-evaluation, and reproducibility of human evaluations.

4. Results

In the original experiment the authors made the following observations with respect to how the neu-

[3]Prolific - https://www.prolific.com

[4]UK Living Wage - https://www.livingwage.org.uk

[5]NeuralREG - https://github.com/ThiagoCF05/NeuralREG

[6]ReproNLP 2024 HEDS forms - https://github.com/nlp-heds/repronlp2024

	Original	Reproduction	CV*
OnlyNames	4.90	4.92	0.4061121348816013
Ferreira	4.93	4.69	4.974662575306527
NeuralREG+Seq2Seq	4.97	4.97	0.0
NeuralREG+CAtt	5.26	4.97	5.652620418943544
NeuralREG+HierAtt	5.13	5.04	1.7646111347510636
Original	5.42	5.22	3.7481401922344113

Table 2: Clarity mean average results from both original and reproduction human evaluation. Unbiased coefficient of variation values (**CV***) calculated using the definition by Belz (2022). Original results are from (Castro Ferreira et al., 2018).

ral models performed against the baseline models and the original texts:

1. *"...all three neural models scored higher than the baselines on all metrics, with especially NeuralREG+CAtt approaching the ratings for the original sentences."*
2. *"...differences between the neural models were small"*
3. *"The results for the 3 different decoding methods of NeuralREG also did not reveal a significant difference."*
4. *"...the original texts were rated significantly higher than both baselines in terms of the three metrics...and than NeuralREG+Seq2Seq in terms of clarity."*

From the results of the reproduction the claims made by the original authors do all hold up and are backed by the results as shown in table 2. This table also includes a column for coefficient of variation for small sample sizes using the methodological approach defined by Belz (2022). Correlations between the original and the reproduction results using both Pearson's *r* of *r=0.783* (*p=0.065*) and Spearman's ρ of ρ*=0.840* (*p=0.036*) were calculated, with both showing statistically significant positive correlations.

With the exception of the *OnlyNames* (slight improvement over original) and the NeuralREG+Seq2Seq systems (same result as original) all other variants showed a decrease in average clarity ratings as compared to the original evaluation. One interesting result is that of the NeuralREG+CAtt system, which showed a marked decrease. Nevertheless, the system still performed as equally as well as the NeuralREG+Seq2Seq and better than the baseline non-neural systems. One possible explanation for the the observed differences could be due to the different cohort of evaluators in the reproduction as compared the original study. The evaluators in the reproduction are much younger and have a greater degree of English language proficiency and this may have lead to the observed variances seen.

5. Conclusion

In this paper we have conducted a successful reproduction of the results obtained in the original human evaluation by Castro Ferreira et al. (2018). There was slight variances in the reported scores in the reproduction, which for a majority of them had slightly lower scores that those originally reported with the exception of two of variants. However, the finding by that the original authors that the neural systems outperform the baselines, whilst underperforming the original text variant holds true and is reconfirmed in this reproduction. In addition, the results obtained in the reproduction show statistically significant positive correlations against the original results.

There are several factors that may have led to this reproduction to having a successful outcome. Factors such as the completeness of the web interface code, the presences of both original collected dataset, and the presence of functional analysis code. Nevertheless, there are areas of improvements. Such as having complete documentation for setting up the experiment. For example, the issue with respect to the order items for each of the six lists could have been mitigated with documentation by the original study authors on the places to look when trying to recreate a given study. Additionally, better documentation would help to remove uncertainty in two aspects. The controls applied for recruiting participants (if any) and the versions of software and libraries used for both the web interface and analysis code. Finally, the missing database schema could of hindered the reproduction experiment from being run at all, but thankfully was worked around with some reverse engineering of the web interface code. Incorporate these improvements would would not only reduce uncertainty, but also reduce the friction in trying to attempt a reproduction by future prospective reproducing authors.

Acknowledgements

Many thanks to *Thomas Khalil* of trivago for his technical support for enabling this reproduction.

6. Bibliographical References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th conference of the European chapter of the association for computational linguistics*, pages 313–320.

Anja Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021a. 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.

Anja Belz and Craig Thomson. 2023. The 2023 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48.

Anya Belz. 2022. A Metrological Perspective on Reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021b. 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*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Maja Popović, and Ehud Reiter. 2022. The 2022 ReproGen shared task on reproducibility of evaluations in NLG: Overview and results. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 43–51, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. The 2024 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

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

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*, pages 1–11.