ReproHum #0712-01: Human Evaluation Reproduction Report for "Hierarchical Sketch Induction for Paraphrase Generation"

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

Abstract

Human evaluations are indispensable in the development of NLP systems because they provide direct insights into how effectively these systems meet real-world needs and expectations. Ensuring the reproducibility of these evaluations is vital for maintaining credibility in natural language processing research. This paper presents our reproduction of the human evaluation experiments conducted by Hosking et al. (2022) for their paraphrase generation approach. Through careful replication we found that our results closely align with those in the original study, indicating a high degree of reproducibility.

Keywords: reproducibility, human evaluation, open science, paraphrase generation

1. Introduction

Human evaluation serves as the cornerstone for appraising the efficacy of machine learning and natural language processing pipelines. Consequently, understanding and addressing the challenges (Howcroft et al., 2020) that may impede the reproducibility of human evaluation experiments is paramount. The ReproHum Project (Belz and Thomson, 2024) is dedicated to devising a methodological framework specifically tailored to assess the reproducibility of human evaluation experiments within the domain of Natural Language Processing (NLP). In line with analogous meta-analytical endeavors (Open Science Collaboration, 2015; Errington et al., 2021a,b), this project seeks to heighten rigor, transparency, and reliability in NLP research. Furthermore, insights garnered from this initiative may help refine future human evaluation methodologies, enhancing their dependability and credibility.

ReproHum has been broken into multiple stages or *rounds*; we are presently near the end of round one. The primary objective of round one is to identify a set of experiments that are reproducible under the same conditions. Additional details regarding round one can be found in §2 and the process is also extensively reported by Belz et al. (2023). Work reported in this paper is part of the second batch of experiments selected for round one of ReproHum.

Specifically, we reproduced the human evaluation experiments conducted in the paper "Hierarchical Sketch Induction for Paraphrase Generation" by Hosking et al. (2022). The original study compared four models for paraphrase generation, and human evaluators assessed the quality of the generated paraphrases. Thanks to the cooperation of the original authors, we were able to reproduce the human evaluation experiments as closely as possible to the original study. We compared our results to the original outcomes, finding that the results of our reproduction are very close to the originally reported results. This suggests that the human evaluation experiments conducted in the original study have a high degree of reproducibility. We have released the data, code, and results of our reproduction to ensure transparency and facilitate further research in this area.[1]

2. Background

In the first step of the ReproHum Project (Belz et al., 2023), 177 papers were identified that (a) contained human evaluation, and (b) were published in the *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* or the *Transactions of the Association for Computational Linguistics (TACL)* in the 2018-2022 period. Through a multi-stage process, 20 experiments from 15 papers were selected to be reproduced. The selection process involved manual review for suitability, responsiveness of the original authors, and availability of a predetermined set of relevant details. Selected experiments were annotated with categorical labels indicating the number of evaluators (small, not small), cognitive complexity[2] (low, medium, high), and evaluator training and expertise (neither, either, both). Following annotation, six of the 20 selected experiments were chosen to achieve a balance of inclusion of these factors in the first batch of reproductions.

In **round one, batch a**, each included experiment was assigned to two partner labs. The partner labs were instructed to reproduce the experiment as closely as possible, given the information

[1] https://github.com/mo-arvan/paraphrase-generation-reproduction

[2] Based on scores given to each criterion in Appendix E of Howcroft et al. (2020).

Proceedings of the HumEval @LREC-COLING 2024, pages 210–220
Turin, Italy. May 21st, 2024.

provided in the original paper and any additional information and clarification obtained through direct communication between the original authors and the ReproHum leadership team. Partner labs were also instructed to document any deviations from the original experiment and the reasons for these deviations. The results of the reproduction were compared to the originally published results to assess the extent to which the experiment was reproducible. These reports and the corresponding data were published as part of the *2023 ReproNLP Shared Task on Reproducibility of Evaluations in NLP* (Belz and Thomson, 2023; González Corbelle et al., 2023; Watson and Gkatzia, 2023; Arvan and Parde, 2023; van Miltenburg et al., 2023; Ito et al., 2023; Gao et al., 2023; Mieskes and Benz, 2023; Hürlimann and Cieliebak, 2023; Platek et al., 2023; Klubička and Kelleher, 2023; Li et al., 2023; Mahamood, 2023).

Overall, the results from round one suggest a varied degree of reproducibility across the experiments, with some being easily reproduced and others not. By analyzing the attributes of each experiment and the corresponding results of the reproduction, it can be inferred that the higher the cognitive complexity, the lower the degree of reproducibility. The total number of evaluators also had an inverse correlation with the degree of reproducibility. While these preliminary findings are insightful, the ReproHum team acknowledged that they are based on a small sample size and may not be generalizable. Hence, another batch of experiments was selected for additional reproducibility assessment (**round one, batch b**). It is our round one, batch b results that we report in this paper.

3. Methods

For round one, batch b, we were assigned to reproduce "Hierarchical Sketch Induction for Paraphrase Generation" (Hosking et al., 2022). The ReproHum leadership team shared a document containing general instructions for reproduction and experiment-specific information for this paper. We summarize the paper and our methods for reproducing it below.

3.1. Hierarchical Sketch Induction for Paraphrase Generation

Hosking et al. (2022) introduced a new generative model called Hierarchical Refinement Quantized Variational Autoencoders (HRQ-VAE). Their proposed model utilized syntactic sketch for paraphrase generation, drawing parallels to the way humans plan out utterances and using those similarities in a sketching step added to the model to help in generating paraphrases. They evaluated the performance of their model compared to several baseline models on the Paralex (Fader et al., 2013), Quora Question Pairs (QQP)[3] and MSCOCO datasets (Lin et al., 2014).

For baselines, the authors compared their approach to Gaussian Variational AutoEncoder (VAE) (Bowman et al., 2016), Latent Bag-of-Words (Fu et al., 2019), Separator (Hosking and Lapata, 2021), and several other paraphrase generation systems. They evaluated their approach and the baselines on the mentioned datasets using iBLEU (Sun and Zhou, 2012), BLEU, Self-BLEU, and P-BLEU. iBLEU, the primary evaluation metric, is a variant of BLEU that uses a paraphrase dataset to evaluate paraphrase quality by assessing the faithfulness of generated outputs compared to ref-

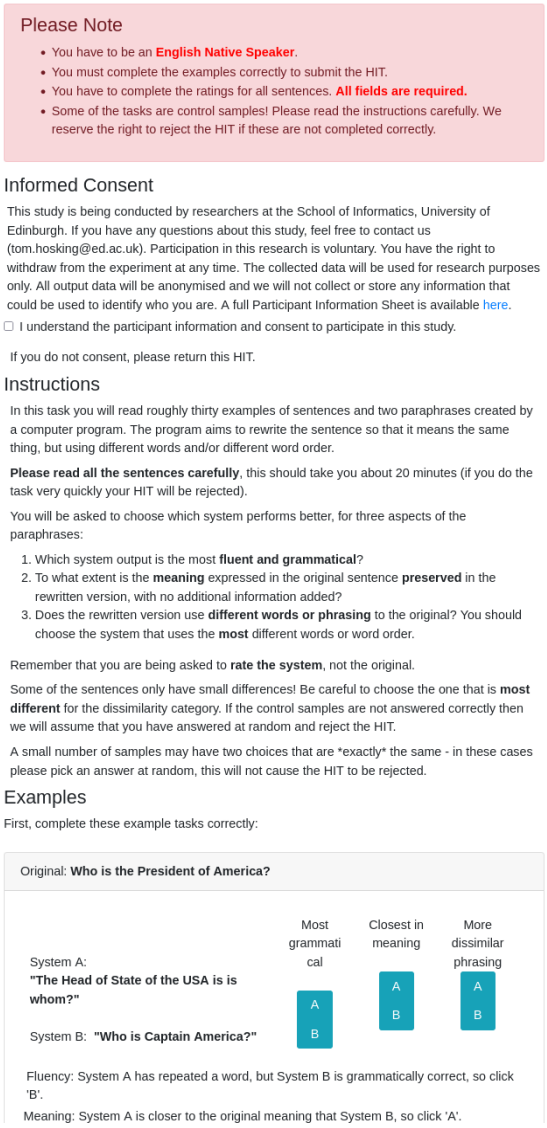

Figure 1: The user interface used for human evaluation in the original study.

[3] https://kaggle.com/competitions/quora-question-pairs

erence paraphrases. It also gauges the extent to which diversity is incorporated. Their automated evaluation suggests that the VAE, Latent BoW, Separator, and HRQ-VAE models performed best.

The four top-performing models were then selected for additional human evaluation. The human evaluation was conducted on Amazon Mechanical Turk (MTurk). It involved 180 human intelligence tasks (HITs), each containing 32 paraphrase pairs.[4] Each task contained two attention checks to ensure the quality of the responses. MTurk workers were asked to select the best paraphrase given an input text and the output of two models, based on three criteria: fluency, meaning, and dissimilarity. Figure 1 shows the user interface used for the human evaluation in the original study. The verbatim instructions of the task included in the user interface are provided below:

- *Which system output is the most **fluent and grammatical**?*
- *To what extent is the **meaning** expressed in the original sentence **preserved** in the rewritten version, with no additional information added?*
- *Does the rewritten version use **different words or phrasing** to the original? You should choose the system that uses the most different words or word order.*

The authors provided additional information regarding the human evaluation in the appendix of their paper. Importantly, they reported utilizing MTurk's feature to make HITs available only in specific regions, setting their region availability to the United States and the United Kingdom. Furthermore, they reported that participants were compensated for their time at a rate above the living wage in the regions selected.

Ultimately, in comparing paraphrase pairs the authors evaluated 300 sentences sampled equally from the three datasets, with paraphrases generated by each model resulting in a total of 1800 paraphrases.[5] For a particular pair of two system outputs for a given input sentence, separately for each of the three criteria, a given system received +1 or -1 depending on whether it was chosen as the best (+1) or worst (-1). The final scores for each model were then calculated by averaging the scores across all of that model's scored samples for a particular criterion. This scoring process is known as Best-Worst Scaling (Louviere and Woodworth, 1991; Louviere et al., 2015).

According to the authors, HRQ-VAE was found to be *more fluent* and *more diverse* while maintaining a *similar meaning* to the original sentence. Figure 4 in their paper shows the results of the human evaluation. We identified five unique **claims** based on the human evaluation results in the original paper:

- **Claim 1:** The VAE baseline is the best at preserving meaning.
- **Claim 2:** The VAE baseline is the worst at introducing variation to the output.
- **Claim 3:** HRQ-VAE better preserves the original intent compared to the other systems.
- **Claim 4:** HRQ-VAE introduces more diversity than VAE.
- **Claim 5:** HRQ-VAE generates much more fluent output than VAE.

3.2. Scope of Reproduction

Our goal was to repeat the allocated experiment as closely as possible to the original study. We focused on a narrow scope of the original paper: we sought to reproduce the outcomes of the human evaluation experiments for the *meaning* criterion. We set up the experiment using all information available to us from the original paper (Hosking et al., 2022) and from follow-up communications with the authors by the ReproHum leadership team. We filled Human Evaluation Datasheet (HEDS) (Shimorina and Belz, 2022) containing the details of the human evaluation experiment. The HEDS is released in the ReproNLP central GitHub repository for HEDS documents.[6]

3.3. Additional Information Obtained from the Original Authors

While we did not directly communicate with the original authors, the ReproHum team provided us with additional information obtained from them. Specifically, the authors shared the exact outputs that they evaluated and the user interface that they used for the human evaluation. Crucially, the authors noted that they used *attention checks* (control samples with known labels). Each task contained two control samples; in one control sample, the system was a "distractor" and the output was a random sample with a completely different meaning that should clearly never be chosen as best for the *meaning* criterion. The other control sample was

[4] Note that *HIT* is a term used on MTurk to refer to a single task or job that a worker can complete; we use the terms *task* and HIT interchangeably in this paper.

[5] There were four systems; for each comparison, we selected two out of four: $\binom{4}{2} = 6$. With the resulting six unique comparisons for each of the 300 sentences, we have a total of $6 \times 300 = 1800$ comparisons.

[6] https://github.com/nlp-heds/repronlp2024

when the system's output was the same as the input, which should clearly never be chosen as best for the *dissimilarity* criterion. Note that the second control sample was not relevant to our reproduction, as we were reproducing the results for the *meaning* criterion. In their communication, the authors mentioned that HITs for which either of these attention checks were failed were rejected and resubmitted to MTurk. Additionally, they reported compensating participants with $3.50 per HIT with an expected completion time of 20 minutes.

3.4. Notes on Experimental Design

The original design of the human evaluation did not consider cases in which both outputs were equally good (a tie). Although we would have preferred to include this option, we followed the original design. Moreover, in analyzing the outputs we uncovered a slight imbalance in the number of samples selected from each dataset. Specifically, while QQP had 100 samples, MSCOCO had 102 samples and Paralex had 98 samples.

3.5. Known Deviations from the Original Experiment

We are aware of several deviations in our reproduction from the original experiment, and we detail these below. We do not believe that these deviations had a major impact on our reproduction results.

Crowdsourcing Platform: Our biggest deviation from the original experiment was in the crowdsourcing platform used. While the original study had utilized MTurk, we used Prolific.[7] This decision was made across all experiments in the ReproHum project to ensure consistency, due to limitations in credit usage on MTurk and the administrative overhead of managing the funds for different experiments.

Prolific survey design is different from MTurk, and we had to adapt the original survey design to the Prolific platform. To be more specific, setting up a survey similar to the structure of HITs was only possible using external survey tools. The ReproHum team shared the code for hosting a server to run the survey. We used a modified version of the code with additional checks to ensure the validity of the responses. Furthermore, we added thread safety to prevent race conditions, where two or more threads try to access or modify the same data at the same time, leading to unpredictable or incorrect results.

Region Control: Our reproduction also deviated slightly in terms of participant region control. While the original authors had limited their HIT availability regions to the United States and the United Kingdom, we followed the region control guidelines of all experiments in the ReproHum project. This meant that participants from Australia and Canada were also included in addition to the United States and the United Kingdom.

Participant Selection: The authors reported filtering participants with approval rates less than 96%, and required that participants had completed at least 5000 HITs. In contrast, we set the approval rate to 99% and the minimum number of HITs completed to 200. This decision was based on the recommendations from Prolific to ensure high-quality participants.[8]

Failed Attention Checks: The original authors reported rejecting HITs for which the attention checks were failed. We did not reject any HITs based on attention checks per recommendations from the ReproHum team; however, we solicited new responses for tasks that failed attention checks.

Participation Limit: The original paper did not report whether a participant could respond to multiple HITs; we assume that no controls were in place for this. In Prolific, participants cannot respond to the same study more than once, even though the input data may be different.

Expected Completion Time: The original authors reported that the expected completion time for a HIT was 20 minutes. Our survey differed from the original study since we only collected responses for the *meaning* criterion. We ran several surveys to estimate the time it would take to complete the task. Ultimately, we set the expected completion time to 8 minutes.

Payment: The original authors reported compensating participants with $3.50 per HIT for 20 minutes of work, resulting in an hourly rate of $10.50. We followed the guidelines of the ReproHum project, setting the wage as the minimum living wage in the United Kingdom (which was higher than our local minimum wage). At the time of data collection, this value was £12 which was equivalent to $15.14 using the exchange rate between UK and US currency at that time. To be more specific,

[7] https://www.prolific.com/

[8] https://www.prolific.com/resources/find-filter-favourite-how-to-select-participants-for-ai-tasks

the participants received £1.60 or $2 for 8 minutes of work.

User Interface: Our institutional consent forms were required to be much more detailed than those used in the original study, and this was beyond our control. To ensure that the participants were not overwhelmed, we split the welcome, instructions, and task into three separate pages. We have included images of the user interface used for the reproduction in the appendix (Figures 3, 4, and 5).

Data Analysis: The source code for the data analysis was intentionally left out and we were asked to write our own code to analyze the data. We also conducted additional analyses to better understand the data. We report our findings from these analyses in §5.

4. Quantified Reproducibility Assessment

We followed the standardized procedure for reproducibility assessment as outlined by the ReproHum team. For single numerical result scores, we calculated the coefficient of variation (CV) to quantify the precision of the results. The CV is calculated as the ratio of the standard deviation of the results to their mean. It serves as a measure of relative variability, and it is useful for comparing the precision of different experiments. We adjusted the CV for small sample sizes as reported by Belz (2022), and refer to this adjusted CV using the notation *CV**. Furthermore, the results are shifted by 100 to ensure the mean is positive, as the original scores were in the range of -100 to 100.

For sets of numerical scores, we calculated Pearson and Spearman correlations between the reproduced and original results. The Pearson correlation measures the linear relationship between two sets of scores, and the Spearman correlation measures the monotonic relationship between two sets of scores. Both correlations range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation (suggesting that the outcomes are diametrically opposed), and 0 indicates no correlation. Using these metrics, we assessed how closely the reproduced results aligned with the original results.

5. Results

5.1. Study Analysis

According to the summary statistics provided by Prolific, the median time spent on the survey was 7 minutes and 12 seconds. With this time, the actual hourly rate was calculated to be £13.30. With filters set for region control and acceptance rate, 51,430 of 152,649 possible participants were eligible to participate in the study; our 180 participants were selected from this pool. We had to repeat one task due to failed attention checks, making the total number of participants n=181.

System	Orig	Ours	CV*	*r*	ρ
VAE	36	37.04	0.76	0.99	1
Latent BoW	-16	-14.52	1.74		
Separator	-24	-29.78	7.88		
HRQ-VAE	4	7.26	3.08		

Table 1: Overview comparing the original and reproduced versions of the human evaluation, including precision metrics to reflect the degree of reproducibility. Pearson's correlation is represented by r and Spearman's correlation is represented by ρ. CV* is computed using n=2. *Orig* refers to the original results reported by Hosking et al. (2022).

Sys.	Win #	Loss #	Best-Worst Score	Best-Worst Scale	Win %
VAE	1850	850	1000	37.04	68.52
Lat. BoW	1154	1546	-392	-14.52	42.74
Sep.	948	1752	-804	-29.78	35.11
HRQ-VAE	1448	1252	196	7.26	53.63

Table 2: Additional details from our own reproduced human evaluation. *Lat. BoW* refers to the *Latent Bag-of-Words* system, and *Sep.* refers to the *Separator* system.

Aside from data available in Prolific, we collected additional data from the survey. Particularly, we collected the time spent on each page of the survey. We present the histogram of time spent on each page of the survey in Figure 6 (in Appendix B). Furthermore, we present the empirical cumulative distribution function (eCDF) of the time spent on each page of the survey in Figure 7 (in Appendix B). Note that this data may not be entirely reliable as participants were given an hour to complete the survey, and the time spent on each page was not necessarily indicative of the time spent on the task (e.g., participants may have stepped away from the computer while leaving the page open). Nonetheless, we consider it a reasonable proxy for the time spent on the task.

The 50th percentiles (median) of the time spent on the welcome, instructions, and task pages were 13, 53, and 328 seconds, respectively. Additionally, we observed that the 90th percentiles of the time spent on the welcome and instruction pages were

82 and 92 seconds, respectively. In other words, the eCDF suggests that 90% of the participants spent less than 82 seconds on the welcome page and 92 seconds on the instruction page. The task page eCDF suggests that the 80% percentile of time spent on the task page was 434 seconds, meaning that 80% of the participants spent less than 434 seconds on the task page. Recall that the total time allotted for the survey was 480 seconds (8 minutes).

5.2. Reproduction Results

Table 1 shows the results of the human evaluation for the selected criterion, comparing the outcomes from the original and reproduced experiments. Overall, we observe that our results are very close to the scores originally reported (Hosking et al., 2022). This is reflected in low CV* values for all the systems. Pearson correlation and p-value are r=0.99 and p=0.01, respectively. Similarly, Spearman correlation and p-value are ρ=1.00 and p=0.00. Both Pearson and Spearman correlations are very high, indicating a strong relationship between the original and reproduced scores. Figure 2 presents this same information in the format used by the original paper, showing best-worst scaling outcomes for the four systems compared in the original paper and in our reproduction.

In Table 2, we include additional details from our own reproduced human evaluation. We report the number of wins and losses for each system, the best-worst score outcome (the sum of all scores of +1 or -1 that the system received), and the best-worst scale outcome. We also report the percentage of wins for each system. We used Krippendorff's alpha to evaluate the agreement among the categorical responses collected, resulting in a value of α=0.51. This metric was not included in the original study, preventing a direct comparison of our findings.

For statistical analysis, we employed ANOVA to determine significant differences among the means of multiple independent groups. We measured effect size using partial eta squared (η^2), which yielded a large effect size of 0.17 for the ANOVA test. With a sample size of 300 and α=0.05, the calculated test power was 0.67, falling below the recommended threshold of 0.80. Achieving a power of 0.80 would require a sample size of 395. In conducting the ANOVA test, we observed an F value of 79.93 with a corresponding p=3.97e-47. Subsequently, we used Tukey's HSD test to identify significant differences between individual groups, revealing significant distinctions among all groups.

Overall, given our reproduced results' similarity to and correlation with the originally reported results, we could easily confirm two out of five of the original claims based on the human evaluation results. The other three claims were out of scope for our reproduction, as they pertained to criteria other than *meaning*. We summarize the claims and our verification in Table 3.

Claim	**Verification**
The VAE baseline is the best at preserving meaning.	Verified
The VAE baseline is the worst at introducing variation to the output.	Out of Scope
HRQ-VAE better preserves the original intent compared to the other systems.	Verified
HRQ-VAE introduces more diversity than VAE.	Out of Scope
HRQ-VAE generates much more fluent output than VAE.	Out of Scope

Table 3: Claims and verifications.

6. Discussion

Since we found this experiment to be underpowered, combining the data collected in our reproduction with the parallel work of the ReproHum project could provide a more robust analysis. This would allow us to draw more reliable conclusions about the reproducibility of the original study. Nonetheless, if the results of the other reproduction are consistent with ours, we believe this experiment is a good candidate for the next round of the ReproHum project, where some variations could be introduced to further investigate the reproducibility of the original study. Replacing the attention check that is not relevant to the *meaning* criterion with a more relevant one could be a good starting point.

We followed the Prolific recommended guidelines for selecting participants, setting the approval rate to 99% and the minimum number of accepted tasks to 200. The problem with this approach is that the number of accepted tasks inflates over time. A better alternate approach would be to select the top k% of workers based on the total number of accepted tasks. A similar concern was raised by González Corbelle et al. (2023). Considering that data collection is essential to machine learning and NLP research, it is important to ensure the quality of the data collected. Lastly, we observed that some work was submitted in timezones other than those associated with the regions selected. This could be due to participants using VPNs or other methods to change their location. In general, timezones are not reliable and can be easily changed. Thus, this is a complex issue that re-

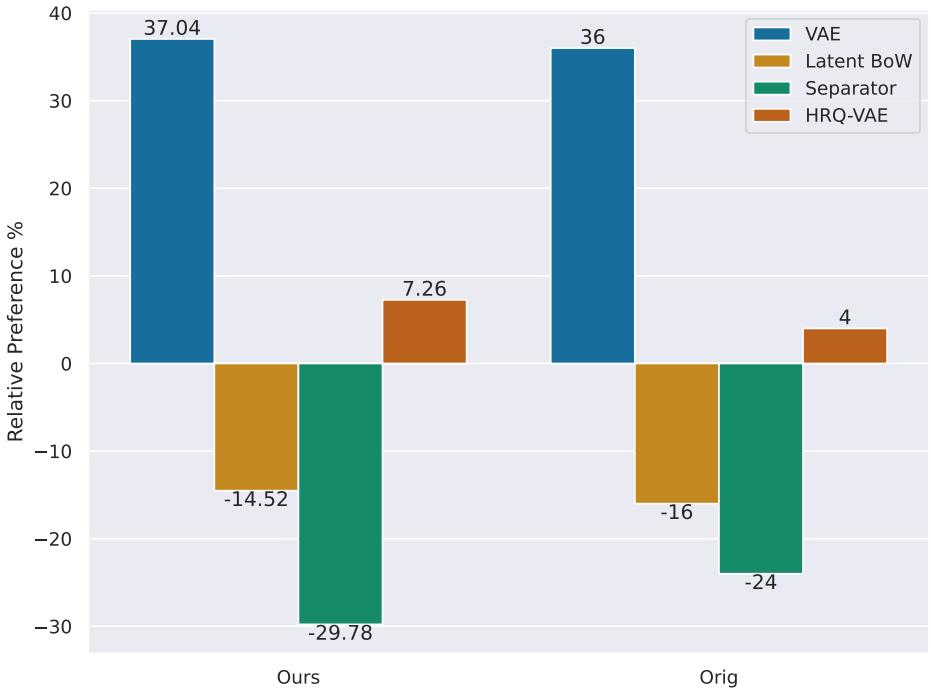

Figure 2: Results of the human evaluation, comparing the original and reproduced systems. Results are presented in the same format used in the original paper.

quires action, cooperation, and transparency from crowdsourcing platforms to ensure the quality of the data collected.

Finally, Platek et al. (2023) report having difficulties setting up the user interface for their reproduction. They suggest utilizing a Docker image containing all the dependencies. We believe that this is a good practice. Considering that our server setup for ReproHum reproductions is customized and unique, we have included the docker compose configuration to bring up the server with all the dependencies and tasks in a separate repository.[9]

7. Conclusion

In this reproduction, we studied the extent to which the human evaluation reported in "Hierarchical Sketch Induction for Paraphrase Generation" is reproducible, narrowing our scope to a single evaluation criterion (*meaning*). We systematically and carefully reproduced the experiment as reported in the original paper to ensure consistency with the original settings to the extent possible. Through a comparison of our reproduced results with those achieved in the original paper using CV*, Pearson's correlation, and Spearman's correlation, we believe that the human evaluation conducted by the original authors has a high degree of reproducibility. This reflects the quality of the design of the experiment. This work would not have been possible without the support of the ReproHum project and the original authors. We hope that our work will contribute to ongoing efforts to improve the reproducibility of research in the field of NLP.

[9] https://github.com/mo-arvan/reprohum-prolific-webapp

Acknowledgments

We would like to thank the ReproHum project, especially Craig Thomson for their support and guidance throughout this reproduction. We would also like to thank the original authors for providing additional information and clarifications. This work was supported by the EPSRC grant EP/V05645X/1.

References

Mohammad Arvan and Natalie Parde. 2023. Human evaluation reproduction report for data-to-text generation with macro planning. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 89–96, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anya Belz. 2022. A metrological perspective on reproducibility in NLP. *Comput. Linguistics*, 48(4):1125–1135.

Anya 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, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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, 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.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 10–21. ACL.

Timothy M Errington, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021a. Challenges for assessing replicability in preclinical cancer biology. *elife*, 10:e67995.

Timothy M Errington, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021b. Investigating the replicability of preclinical cancer biology. *Elife*, 10:e71601.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1608–1618. The Association for Computer Linguistics.

Yao Fu, Yansong Feng, and John P. Cunningham. 2019. Paraphrase generation with latent bag of words. 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 13623–13634.

Mingqi Gao, Jie Ruan, and Xiaojun Wan. 2023. A reproduction study of the human evaluation of role-oriented dialogue summarization models. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 124–129, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Javier González Corbelle, Jose Alonso, and Alberto Bugarín-Diz. 2023. Some lessons learned reproducing human evaluation of a data-to-text system. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 49–68, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tom Hosking and Mirella Lapata. 2021. Factorising meaning and form for intent-preserving paraphrasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1405–1418. Association for Computational Linguistics.

Tom Hosking, Hao Tang, and Mirella Lapata. 2022. Hierarchical sketch induction for paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2489–2501. Association for Computational Linguistics.

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. 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, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 169–182. Association for Computational Linguistics.

Manuela Hürlimann and Mark Cieliebak. 2023. Reproducing a comparative evaluation of German text-to-speech systems. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 136–144, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Takumi Ito, Qixiang Fang, Pablo Mosteiro, Albert Gatt, and Kees van Deemter. 2023. Challenges in reproducing human evaluation results for role-oriented dialogue summarization. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 97–123, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Filip Klubička and John D. Kelleher. 2023. HumEval'23 reproduction report for paper 0040: Human evaluation of automatically detected over- and undertranslations. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 153–189, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yiru Li, Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Same trends, different answers: Insights from a replication study of human plausibility judgments on narrative continuations. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 190–203, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

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.

Saad Mahamood. 2023. Reproduction of human evaluations in: "it's not rocket science: Interpreting figurative language in narratives". In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 204–209, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Margot Mieskes and Jacob Georg Benz. 2023. h_da@ReproHumn – reproduction of human evaluation and technical pipeline. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 130–135, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Ondrej Platek, Mateusz Lango, and Ondrej Dusek. 2023. With a little help from the authors: Reproducing human evaluation of an MT error detector. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 145–152, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 38–42. The Association for Computer Linguistics.

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, and Emiel Krahmer. 2023. How reproducible is best-worst scaling for human evaluation? a reproduction of 'data-to-text generation with macro planning'. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 75–88, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Lewis Watson and Dimitra Gkatzia. 2023. Unveiling NLG human-evaluation reproducibility: Lessons learned and key insights from participating in the ReproNLP challenge. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 69–74, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

A. Reproduction User Interface

We show our reproduced interface for the human subject consent page for the human evaluation in Figure 3. Participants were required to consent by clicking the "Accept & Continue" link prior to taking part in the evaluation. In Figures 4 and 5 we present the participant and template views for the reproduced evaluation, respectively.

B. Time Spent on Survey

In Figures 6 and 7 we report the amount of time spent by participants on the reproduced evaluation. Time was recorded for each page of the survey. Figure 6 shows a histogram of the number

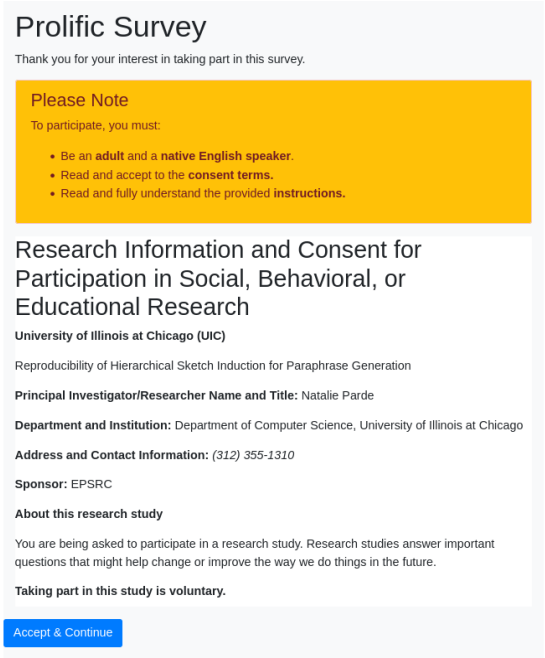

Figure 3: Reproduced interface for the human evaluation (consent page).

of seconds spent on each page, whereas Figure 7 computes and displays an empirical cumulative distribution function for this data.

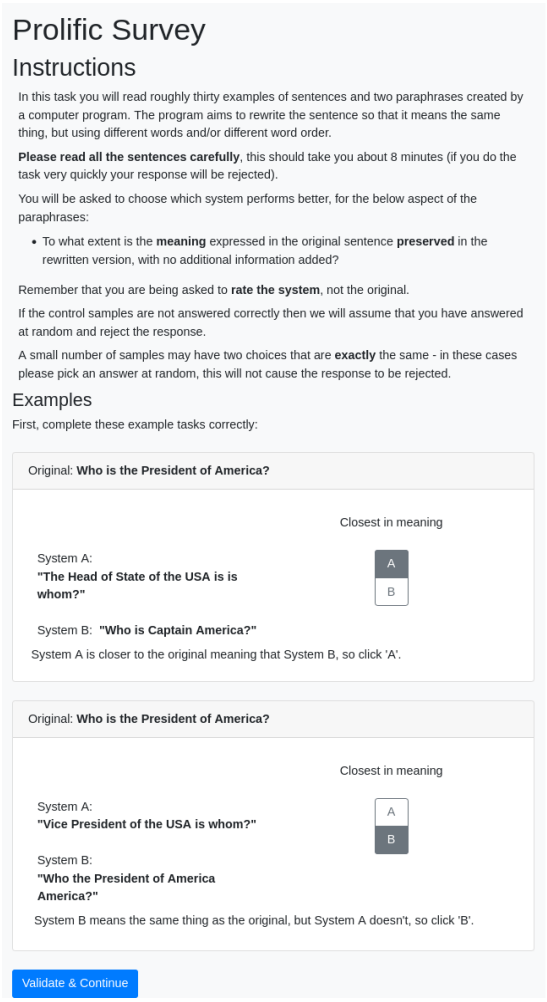

Figure 4: Reproduced interface for the human evaluation (participant view).

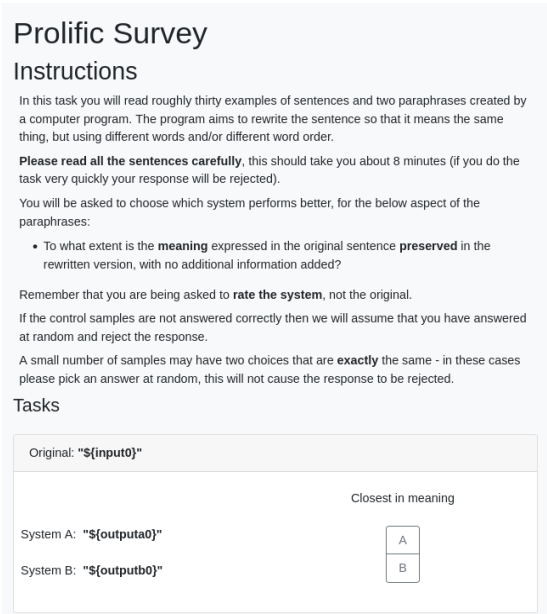

Figure 5: Reproduced interface for the human evaluation (template view).

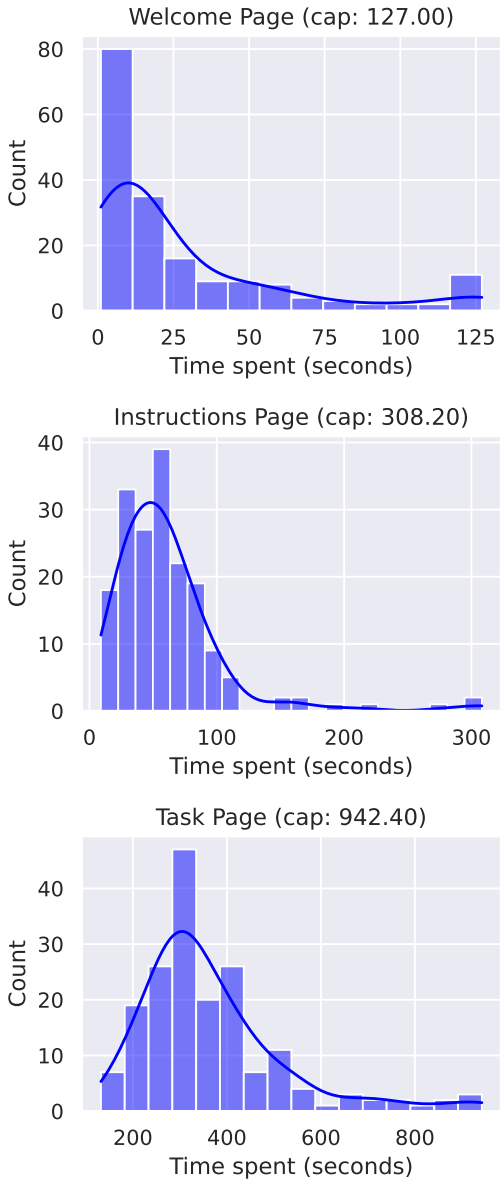

Figure 6: Histogram of seconds spent on each page of the survey. Note that each histogram is capped to ensure readability.

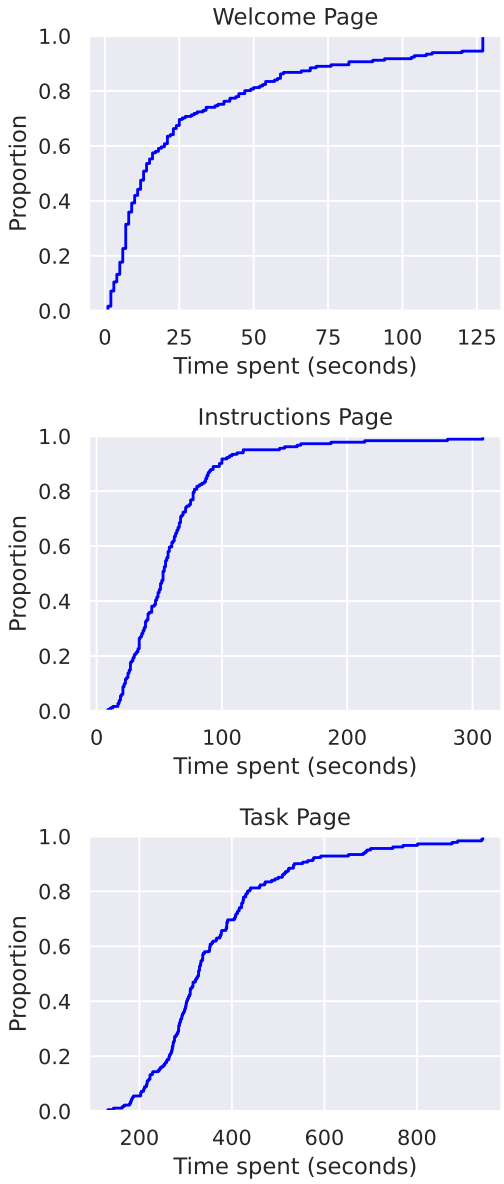

Figure 7: Empirical Cumulative Distribution Function (eCDF) of seconds spent on each page of the survey.