ReproHum #0712-01: Reproducing Human Evaluation of Meaning Preservation in Paraphrase Generation

Lewis Watson, Dimitra Gkatzia
Edinburgh Napier University
{L.Watson, D.Gkatzia}@napier.ac.uk

Abstract

Reproducibility is a cornerstone of scientific research, ensuring the reliability and generalisability of findings. The ReproNLP Shared Task on Reproducibility of Evaluations in NLP aims to assess the reproducibility of human evaluation studies. This paper presents a reproduction study of the human evaluation experiment in "Hierarchical Sketch Induction for Paraphrase Generation" by Hosking et al. (2022). The original study employed a human evaluation on Amazon Mechanical Turk, assessing the quality of paraphrases generated by their proposed model using three criteria: meaning preservation, fluency, and dissimilarity. In our reproduction study, we focus on the meaning preservation criterion and utilise the Prolific platform for participant recruitment, following the ReproNLP challenge's common approach to reproduction. We discuss the methodology, results, and implications of our reproduction study, comparing them to the original findings. Our findings contribute to the understanding of reproducibility in NLP research and highlights the potential impact of platform changes and evaluation criteria on the reproducibility of human evaluation studies.

Keywords: reproducibility, NLG, paraphrase generation, human evaluation

1. Introduction

Reproducibility is a fundamental principle of scientific research, ensuring that findings can be independently verified and built upon by the wider research community. In the field of Natural Language Generation (NLG), reproducibility is particularly challenging due to the complex nature of the tasks (Belz et al., 2023) and the use of human assessments for the evaluation of NLG approaches (Gehrmann et al., 2023; Howcroft et al., 2020). Recently, the reproducibility of NLP studies has been called into question, with concerns raised about the reliability and generalisability of reported findings (Belz et al., 2021).

The ReproNLP Challenge To address the issue of reproducibility in NLP, the ReproNLP/ReproGen challenge was established to assess the reproducibility of human evaluation studies (Belz et al., 2020), under three conditions: (1) reproduction of evaluation results of pre-selected papers based on information of the original paper and additional information by the authors; (2) reproduction of evaluation results by the same authors, i.e. own study; (3) reproduction of a pre-selected study using information provided by the ReproNLP organisers only (Belz and Thomson, 2023).

The 2023 round of reproduction studies provided a wealth of lessons learnt. The evaluators' background and qualifications were identified as important factors in obtaining consistent results as discrepancies in these can lead to varying results. (González Corbelle et al., 2023; Watson and Gkatzia, 2023; Mieskes and Benz, 2023; Li et al., 2023; Mahamood, 2023). The number of ratings obtained per item and worker are also important for obtaining statistically similar results (van Miltenburg et al., 2023; Ito et al., 2023; Gao et al., 2023). In addition, Ito et al. (2023) highlight that errors in statistical analyses can prohibit reproducibility. Technical issues can prohibit replications of studies that can be overcome through the use of Docker (Platek et al., 2023) and the provision of code used for crowdsourcing and the analysis of results (Mahamood, 2023). However, Klubička and Kelleher (2023) used a different user interface for their reproduction study than the original authors and were able to confirm the results of the original study. Discrepancies in the study design have also been identified as an issue in reproducibility (Platek et al., 2023; Gao et al., 2023), while Platek et al. (2023) advocate that setups with a minimal range of potential answers, particularly those with binary questions, are simpler to duplicate and should be favoured over more intricate setups whenever feasible.

In the 2023 round, we reproduced the human evaluation as close as possible to the methodology used by the original authors (Watson and Gkatzia, 2023). In the 2024 round, we experimented with a platform change, Prolific instead of Amazon Mechanical Turk, and we focused on only one quality criterion (meaning preservation) as outlined in the 2024 challenge design (Belz and Thomson, 2024).

Hierarchical Sketch Induction for Paraphrase Generation In this paper, we focus on reproducing a single quality criterion from the human evaluation component from the study "Hierarchical Sketch

Induction for Paraphrase Generation" by Hosking et al. (2022). The original study proposed a novel approach to paraphrase generation using hierarchical sketch induction and conducted a human evaluation on Amazon Mechanical Turk to assess the quality of the generated paraphrases based on three criteria: meaning preservation, fluency, and dissimilarity.

Human Evaluation Datasheet (HEDS) As part of our reproduction study, we have completed the Human Evaluation Datasheet (HEDS) (Shimorina and Belz, 2022), a standardised template for documenting human evaluation experiments in NLP. The HEDS framework aims to promote reproducibility and facilitate meta-evaluation of evaluation methods by providing a consistent format for recording the details of human evaluations. Our completed HEDS document is available on our project's GitHub repository[1] and has also been contributed to the central HEDS repository maintained by the ReproNLP organisers[2]. This central repository serves as a comprehensive resource for HEDS documents from all participating teams, enabling access and comparison of human evaluation methodologies across different studies. By adhering to the HEDS framework and sharing our documentation, we aim to support the broader goal of improving the reliability and generalisability of human evaluation practices in the field.

1.1. Objectives and Hypotheses

The main objective of our reproduction study is to assess the reproducibility of the human evaluation results reported in Hosking et al. (2022). We aim to answer the following research questions:

1. To what extent can the human evaluation results be reproduced using a different participant recruitment platform (Prolific instead of Amazon Mechanical Turk)?
2. How does focusing on a single evaluation criterion (meaning preservation) affect the reproducibility of the results compared to the original study, which used three criteria?

Based on these research questions, we hypothesise that:

1. The change in participant recruitment platform may lead to some differences in the evaluation results, but the overall trends should remain consistent with the original study.
2. Focusing on a single evaluation criterion may result in higher reproducibility compared to the original study, as it reduces the complexity of the task and the potential for variability in participant judgements.

[1] https://github.com/NapierNLP/repronlp_2024
[2] https://github.com/nlp-heds/repronlp2024

2. Original Study

2.1. Methodology

The original study by Hosking et al. (2022) proposed a novel approach to paraphrase generation called Hierarchical Refinement Quantized Variational Autoencoders (HRQ-VAE). The HRQ-VAE model learns to generate paraphrases by first inducing a syntactic sketch of the input sentence, which captures its syntactic structure at varying levels of granularity. The model then generates the final paraphrase based on the induced sketch and the original sentence's meaning representation.

To evaluate the quality of the generated paraphrases, the authors conducted a human evaluation study on Amazon Mechanical Turk (AMT)[3]. The annotators were required to have an approval rate of >96%, be located in the United States or United Kingdom, and have completed >5000 HITs, workers were paid $3.50USD/hr. They compared the HRQ-VAE model's output to paraphrases generated by three other baseline models, namely, Gaussian Variational AutoEncoder (VAE Bowman et al. 2016), Separator (Hosking and Lapata, 2021) and Latent bag-of-words (BoW, Fu et al. 2019)[4].

The human evaluation tasks were created using 300 input sentences sampled equally from three datasets: Paralex (Fader et al., 2013), Quora Question Pairs (QQP) (Chen et al., 2017), and MSCOCO (Lin et al., 2014). For each input sentence, the paraphrases generated by the HRQ-VAE model and the baseline models were presented to the AMT workers, who were asked to rate the paraphrases based on three criteria:

1. **Meaning preservation**: The extent to which the generated paraphrase preserves the meaning of the original input sentence.
2. **Fluency**: The fluency and grammaticality of the generated paraphrase.
3. **Dissimilarity**: The degree to which the generated paraphrase differs from the original input sentence in terms of word choice and sentence structure.

Each comparison was evaluated by 3 distinct AMT workers, resulting in a total of 900 judgements (300 sentences × 3 judgements per sen-

[3] https://www.mturk.com
[4] The authors compared their model to additional models through automatic metrics, but picked these four for human evaluation due to best performance

tence). Each task contained 32 paraphrase questions, including 2 attention checks.

The first attention check focused on the meaning criteria and consisted of comparisons where one paraphrase is generated by a "distractor" model designed to produce output with a completely different meaning. The second attention check focused on the dissimilarity criteria where the paraphrase would be the same as the input. Where a participant failed the attention check, their results were discarded.

2.2. Results

The original study reported the human evaluation results as relative preference scores for each of the three dimensions (meaning, dissimilarity, and fluency) across the four models: HRQ-VAE, Separator, Latent BoW, and VAE. The relative preference scores were calculated by assigning a score of +1 when a system was selected, -1 when the other system was selected, and taking the mean over all samples.

Key findings from the original study include:

- The VAE baseline achieved the highest relative preference score for meaning preservation (+36%) but the lowest for dissimilarity (-33%), indicating that while it best preserved the original sentence's meaning, it introduced the least variation in the generated paraphrases.
- The HRQ-VAE model offered the best balance between meaning preservation (+4%) and dissimilarity (-3%), demonstrating its ability to generate paraphrases that maintain the original meaning while introducing diversity.
- In terms of fluency, the HRQ-VAE model outperformed Separator and Latent BoW, with a relative preference score of +8%.

These findings highlighted the effectiveness of the proposed hierarchical sketch induction approach in generating high-quality paraphrases that strike a balance between meaning preservation and dissimilarity while maintaining fluency.

3. Reproduction

3.1. Methodology

Our reproduction study aims to assess the reproducibility of the human evaluation results reported in the original study by Hosking et al. (2022). We follow the ReproNLP challenge's common approach to reproduction (Belz et al., 2020), with some modifications to the participant recruitment process and the evaluation criteria.

3.1.1. Participant Recruitment

We recruited participants using the Prolific crowdsourcing platform[5], which differs from the original study's use of Amazon Mechanical Turk (AMT). Participants were sourced from the United Kingdom, Canada, the United States, and Australia to ensure a diverse sample and adhere to the ReproNLP Challenge. To prevent overlap with the participant pool of another lab conducting a similar reproduction study, we exclude participants who have taken part in their study. Additionally, in accordance with the ReproHum common procedure for calculating fair pay (Belz et al., 2023), participants were paid £2. This was calculated by assuming the reduced complexity task should take around 10 minutes and paying £12/hr. The median time to complete the task was 8 minutes and our average reward per hour came to £14.75.

In contrast to the original study, we did not impose any restrictions on the participants' approval rate or number of previously completed tasks on Prolific.

3.1.2. Evaluation Tasks and Procedure

We use the same set of 300 sentences as in the original study. These sentences are divided into 60 distinct tasks (each needing three participant ratings, therefore requiring 180 participants), each containing 32 paraphrase questions, including 2 attention checks.

A single question in a task consisted of an original sentence along with two corresponding paraphrases, each generated by distinct models. Contrasting with the methodology of the original study, our reproduction concentrated solely on a singular criterion. This decision was informed by the preliminary ReproHum findings, which indicated that tasks of lower complexity yielded enhanced reproducibility (Belz et al., 2023). The participants' assigned task was to identify the paraphrase that most effectively retained the meaning of the original sentence.

Each distinct task was evaluated by 3 participants, resulting in a total of 180 participant results (60 distinct tasks × 3 participants per task). After removing the attention check questions, we obtain a total of 1,800 final average comparisons (5760 total evaluations ÷ 3 participants = 1920 average from participants, then 1920 - 120 attention checks = 1800 final). The four models being evaluated in this study are the same as in the original study: VAE, Latent BoW, Separator, and HRQ-VAE.

3.2. Attention Check

To ensure the quality of the collected data, we incorporate an attention check mechanism in our

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

reproduction study, following the same approach as the original paper. The attention check consists of comparisons where one of the paraphrases is generated by a "distractor" model, which is designed to produce output with a completely different meaning from the original sentence. If a participant selected the distractor model, their responses were discarded and reran but we did still pay the participant. We had 5 failed attention checks in the initial run of the reproduction, and then a further 1 failed attention checks on the rerun totalling 6 failed attention checks. There are 2 attention checks per task, and with 60 distinct tasks, there are a total of 120 attention check questions (2 x 60 = 120).

We decided to include the original study's second attention check question, to minimise the differences between the original study and the reproduction however, the data was not used for analysis.

3.3. Preference Calculation

To analyse the results of the reproduction study, we follow the same approach as the original study. For each comparison between two paraphrases, we assign a score of +1 to the model whose paraphrase is selected by the participant as better preserving the meaning of the original sentence. Conversely, the model whose paraphrase is not selected receives a score of -1. This scoring method allows us to calculate the relative preference for each model. The analysis is performed using a Python script, which can be found alongside our raw results on GitHub[6]. The script reads the data from a CSV file and iterates over each unique task number (1-60). For each task, it examines the participant responses for the meaning preservation criterion across all 32 comparisons, excluding the attention check questions.

For each comparison, the script determines the preferred model based on the majority vote across the three participants. If model A is preferred, it receives a score of +1, while model B receives a score of -1, and vice versa. These scores are accumulated for each model across all comparisons.

After processing all the comparisons, the script calculates the total number of comparisons (1800 once the attention checks have been removed) and the average number of preferences across all models. Finally, it computes the relative preference percentage for each model by dividing its accumulated score by the total number of comparisons and multiplying by 100.

The resulting relative preference percentages provide insights into the performance of each model in terms of meaning preservation, as judged by the participants in the reproduction study. These percentages can then be compared to the original study's results to assess the reproducibility of the findings.

[6]https://github.com/NapierNLP/repronlp_2024

3.4. Differences from the Original Study

Our reproduction study differs from the original study in the following aspects:

- We use the Prolific platform for participant recruitment instead of Amazon Mechanical Turk.
- We do not impose restrictions on participants' approval rate or number of previously completed tasks.
- We focus on a single evaluation criterion (meaning preservation) instead of three criteria (meaning, dissimilarity, and fluency).
- We recruited participants from the United Kingdom, Canada, the United States, and Australia as opposed to just the UK and USA like the original study.

These differences allow us to investigate the impact of participant recruitment platforms, screening criteria, and evaluation criteria on the reproducibility of the human evaluation results. It is important to note that some of these changes were planned, such as focusing on a single evaluation criterion and recruiting participants from additional countries, while others, like the omission of participant approval rate and task completion restrictions, were unintentional.

The omission of Prolific filters was an oversight, however it highlights the challenges of conducting reproduction studies with complete accuracy. As Thomson et al. (2024) argues, mistakes might occur in many human evaluations, and there is no evidence to suggest that all published studies are entirely mistake-free. Despite our best efforts to adhere to the original study's methodology, this unintended difference in participant screening criteria may have introduced additional variability in our reproduction results.

4. Results

In this section, we present the results of our reproduction study and compare them with the findings of the original study by Hosking et al. (2022).

Figure 1 illustrates the relative preference results from our reproduction study. The HRQ-VAE model achieves a relative preference score of 3.56%, indicating a slight preference for its generated paraphrases in terms of meaning preservation. The VAE model performs the best, with a score of 23.00%, while the Separator and Latent BoW models receive negative scores of -17.89% and -8.67%, respectively.

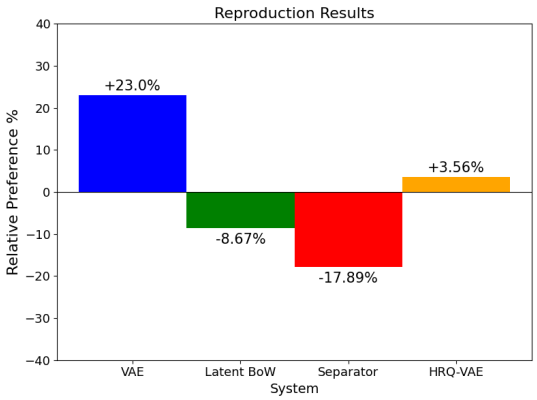

Figure 1: Relative preference results from our reproduction.

To facilitate a direct comparison with the original study, we present the results obtained by Hosking et al. (2022) in Figure 2. The original study reports relative preference scores of +36% for the VAE model, -16% for Latent BoW, -24% for Separator, and +4% for HRQ-VAE.

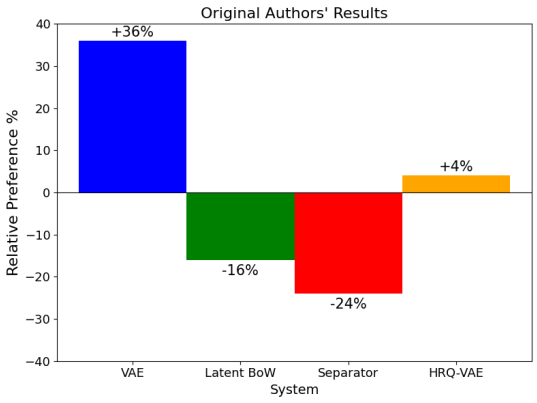

Figure 2: Results obtained by the original authors in Hosking et al. (2022), visualised in a manner consistent with our own findings. The numerical values presented are directly sourced from the original authors' publication.

Comparing the results of our reproduction study with the original findings, we observe some notable differences. While the VAE model maintains its position as the best-performing model in both studies (for preserving meaning), the relative preference scores for the other models vary. In our reproduction, the HRQ-VAE model is slightly less preferred (3.56%) than in the original study (4%). The Separator model is more preferred in our study (-17.89%) compared to the original (-24%), while the Latent BoW model is less preferred (-8.67%) than in the original (-16%), with negative scores indicating less preference for the model. Overall, our replication study shows a narrowing in the range of model preferences: the best models are not as strongly preferred, and the least preferred models are not as strongly disliked as in the original study, even though the ranking order remains the same.

4.1. Quantified Reproducibility Assessments (QRA)

To further evaluate the reproducibility of the original study, we conducted Quantified Reproducibility Assessments (QRA) as described by Belz et al. (2021). These assessments provide a standardised way to quantify the degree of reproducibility between the original study and our reproduction. The code used to do these calculations can be found alongside the data on our github repo[7].

4.1.1. Type I Assessment

Type I assessment measures the reproducibility of individual results using the coefficient of variation (CV*). CV* is an adjusted version of the coefficient of variation that accounts for small sample sizes (Belz, 2022). It can be used even with pairs of results, such as those obtained from an original study and its reproduction. We calculated the CV* for each model by comparing the original and reproduction percentage scores.

$$CV^* = \left(1 + \frac{1}{4n}\right) \frac{s^*}{|\bar{x}|} \quad (1)$$

where s^* is the unbiased sample standard deviation, $\bar{x}$ is the sample mean, and n is the sample size.

Table 1: Type I (CV*) Assessment

System	Original	Reproduction	CV*
VAE	+36%	+23%	43.936
Latent BoW	-16%	-8.67%	59.246
Separator	-24%	-17.89%	29.084
HRQ-VAE	+4%	+3.56%	11.605

4.1.2. Type II Assessment

Type II assessment evaluates the reproducibility of a set of results using correlation measures. We calculated Pearson's r and Spearman's ρ correlations between the original and reproduction percentage scores.

Table 2: Type II (Correlation) Assessment

Metric	Value	p-value
Pearson's r	0.995	0.0049
Spearman's ρ	1.000	<0.0001

[7] https://github.com/NapierNLP/repronlp_2024

5. Discussion

Our reproduction study aimed to assess the reproducibility of the human evaluation results reported by Hosking et al. (2022) for their proposed hierarchical sketch induction approach to paraphrase generation. By closely following their methodology but using the Prolific platform for participant recruitment, only screening participants based on location and focusing on the meaning preservation criterion, we sought to determine to what extent the original findings could be replicated.

The results of our reproduction study show a similar trend to the original findings, with the VAE model clearly achieving the highest relative preference score for meaning preservation. However, we observed some notable differences in the magnitudes of the relative preference scores for the other models. The HRQ-VAE model, which was the main focus of the original study, received a slightly lower preference score in our reproduction (3.56%) compared to the original (4%). Additionally, the Separator and Latent BoW models exhibited different degrees of dislike compared to the original study. The Separator model was less disliked in our reproduction, with a relative preference score of -17.89% compared to -24% in the original study. Similarly, the Latent BoW model was also less disliked in our reproduction, receiving a score of -8.67% compared to -16% in the original study.

To further evaluate the reproducibility of the original study, we conducted Quantified Reproducibility Assessments (QRA) as described by Belz et al. (2021). The assessment of individual model reproducibility using the coefficient of variation (CV*) revealed some variability, with the Separator model showing the best reproducibility (CV* = 29.0843) and the Latent BoW model having the lowest reproducibility (CV* = 59.2464). However, the assessment of the overall reproducibility using correlation measures demonstrated a strong positive correlation between the original and reproduction results. Both Pearson's r (0.995, $p = 0.0049$) and Spearman's ρ (1.000, $p < 0.0001$) indicated a high degree of overall reproducibility.

Despite these differences, the overall ranking of the models in terms of meaning preservation remained consistent between the original study and our reproduction. This suggests that the fundamental findings of the original study are reproducible to some extent, even with the modifications made to the participant recruitment platform, and the focus on a single evaluation criterion.

It is important to acknowledge the limitations of our reproduction study. First, the use of a different participant recruitment platform (Prolific) and the exclusion of certain participant screening criteria may have introduced variability in the evaluator pool, potentially influencing the results. Second, focusing on a single evaluation criterion (meaning preservation) rather than the three criteria used in the original study may have simplified the task for participants but also limited the scope of the reproducibility assessment.

6. Conclusion

Our findings contribute to the broader discussion on the reproducibility of human evaluation studies in NLP research. The fact that we were able to largely reproduce the original results, despite the modifications made, highlights the potential for reproducing human evaluation findings across different platforms and with variations in the evaluation setup. However, the observed differences in the relative preference scores underscore the sensitivity of human evaluations to factors such as participant recruitment and the specific evaluation criteria used.

To further enhance the reproducibility of human evaluation studies, we recommend that researchers provide detailed documentation of their methodology, including participant recruitment procedures, evaluation guidelines, and analysis methodologies. Additionally, we strongly suggest publishing both raw data and analysis code where possible. This transparency will facilitate replication attempts and enable more robust comparisons across studies. Additionally, exploring the impact of different participant pools and evaluation setups on the reproducibility of results can provide valuable insights into the generalisability of human evaluation findings.

In conclusion, our reproduction study demonstrates that the human evaluation results reported by Hosking et al. (2022) are partially reproducible when using a different participant recruitment platform and focusing on a single evaluation criterion. While we observed some differences in the relative preference scores, the overall ranking of the models remained consistent with the original findings. This study contributes to the ongoing efforts to assess and improve the reproducibility of human evaluation studies in NLP research, and highlights the importance of detailed documentation and exploration of factors influencing reproducibility. Future work should continue to investigate the robustness of human evaluation findings across different setups and participant pools to strengthen the reliability and generalisability of NLP evaluation practices.

7. Bibliographical 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.

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. 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*, pages 232–236, Dublin, Ireland. 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*, pages 381–393, Online. Association for Computational Linguistics.

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.

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, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, 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 *Proceedings of 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 Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Zihang Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. Quora question pairs.

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 (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria. Association for Computational Linguistics.

Yao Fu, Yansong Feng, and John P Cunningham. 2019. Paraphrase generation with latent bag of words. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

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.

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 (Volume 1: Long Papers)*, pages 1405–1418, Online. 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)*, pages 2489–2501,

Dublin, Ireland. 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*, pages 169–182, Dublin, Ireland. 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.

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.

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.

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.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

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.

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.

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.

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.

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.

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.