Reproduction of Human Evaluations in: “It’s not Rocket Science: Interpreting Figurative Language in Narratives”

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

Abstract

We describe in this paper an attempt to reproduce some of the human of evaluation results from the paper “It’s not Rocket Science: Interpreting Figurative Language in Narratives”. In particular, we describe the methodology used to reproduce the chosen human evaluation, the challenges faced, and the results that were gathered. We will also make some recommendations on the learnings obtained from this reproduction attempt and what improvements are needed to enable more robust reproductions of future NLP human evaluations.

1 Introduction

Reproducible and repeatable evaluations lay at the heart of good science. However, there has been increasing concern with Natural Language Processing (NLP) on whether human evaluations are in fact reproducible and repeatable. This is particularly important within the field of NLP as human evaluations are seen as the “gold standard” as compared to automatic metric based evaluations. This has lead to an interest in trying to understand and quantify the degree to which evaluations are reproducible.

One such effort is the ReproHum project[1], which attempts to investigate human evaluations within NLP by systematically uncovering the extent of problems of reproducibility. As part of this project multiple partner labs, consisting of both academic and industry institutions, were invited to participate in a multi-lab study reproducing human evaluations from a chosen set of research papers. These papers were vetted by the organising committee of the ReproHum project 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.

In this paper, we describe the current challenges facing human evaluations in NLP and reproducibility (section 2). Afterwards we give details on the attempt to reproduce a specific human evaluation within the paper “It’s not Rocket Science: Interpreting Figurative Language in Narratives” by Chakrabarty et al. (2022) (section 3) and how the reproduction of the paper was conducted with details on the challenges involved (section 4). Finally, we detail the results obtained from the reproduction (section 5) and the recommendations (section 6) we would make based on the experiences of this experiment that would enable more robust reproductions of future NLP human evaluations.

2 Background

Within recent years there has been a great interest in understanding and quantifying the reproducibility of experiments across several areas of scientific research. This also includes experiments in the field of Natural Language Understanding (NLU), where researchers have questioned the degree to which experiments and results can reliably be reproduced. Recent work exploring the reproducibility of past NLU work has found significant issues such as only a minority of systems reproducing previously reported scores and systems not working due to 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). Additionally, there has been growing awareness of systematic issues with regard to how human evaluations are being conducted. In particular, the lack of standardisation and significant underreporting of key human evaluation details (Howcroft et al.,

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

2020). There has been an attempt to make human evaluation reporting more standardised and comparable between different papers through an introduction of a classification system that defines quality criterion properties (Belz et al., 2020b). However, as noted by Gehrmann et al. (2023), whilst the problems of evaluations are known and proposals have been made to improve the situation, the adoption of best practices remains lacking.

The ReproHum project is a subsequent follow-up of the ReproGen shared tasks[2] (Belz et al., 2020a) in 2021 and 2022. In these shared tasks participants either selected a paper proposed by the organisers or self-selected a paper for human evaluation reproduction. The results from these shared tasks showed some indications that human evaluations that have different evaluation cohorts can disadvantage the reproducibility of a given experiment (Belz et al., 2021a). However, lowering the cognitive loads on individual evaluators could potentially lead to be better reproducibility of results (Belz et al., 2022).

3 Reproduction Experiment

For this reproduction experiment we were tasked with reproducing a specific human evaluation in the paper "It's not Rocket Science: Interpreting Figurative Language in Narratives" by Chakrabarty et al. (2022). The paper explores the interpretation of figurative languages (idioms and similes) in English by exploring plausible and implausible continuations from a given fictional narrative. The authors of the paper used models to generate plausible idioms and similes from a given narrative. These generated texts were compared to human written ones in both automatic and human evaluations.

The focus for this experiment is to reproduce the human evaluation conducted by authors. In particular, reproducing the absolute human evaluation, which asked human Amazon Mechanical Turk workers on whether the computer generated and/or the human references are plausible or not for the given narrative. This task is illustrated in figure 1, which is taken from Chakrabarty et al. (2022). In the original experiment crowd workers were shown a narrative, the meaning of the idiom or the property of the simile and a list of three automatically generated continuations. One from a baseline supervised GPT-2 model, one from a context model, and the third from the literal model. In addition to the automatic continuations, participants were shown three human alternatives for idioms or five for similes. For each continuation (automatic or human) participants were asked to rate whether the text is plausible or not. Each example was rated by three workers and the result aggregated using majority voting.

Both evaluations were done on 25 randomly sampled narrative texts for both the absolute idiom and simile scenarios. This equates to 50 narrative texts in total. The original paper incorrectly states "50 narratives for each task", however prior to the reproduction experiment this was clarified by the authors to be a mistake.

In the original evaluation the authors of the paper reported the following results for the absolute evaluation:

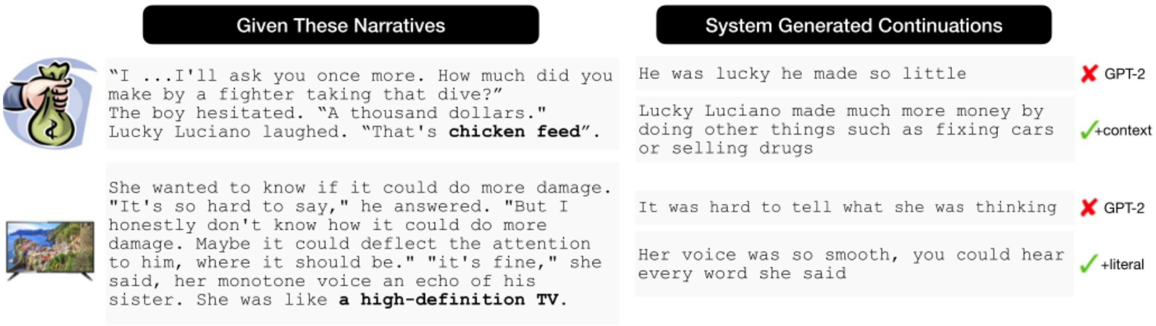
- Moderate inter-annotator agreement using Krippendorf's $\alpha = 0.68$.
- 80% of human-written continuations for the idiom and 88% for simile tasks were judged as plausible.
- 56% of the baseline GPT-2 model continuations for the idiom and 60% for the simile tasks were judged as plausible.
- 68% of the context model continuations for both idiom and simile tasks were judged as plausible.
- 48% of the literal model continuations for the idiom and 76% for the simile tasks were judged as plausible.

In addition to the above reported results the authors also make a mention of the fact that "the context model was favoured for idioms, the literal model was favoured for similes". This result will also be checked in this reproduction attempt.

4 Methodology & Challenges

The original evaluation collected human evaluations using Amazon Mechanical Turk (MTurk) crowd workers. Like the original, the reproduction also used Amazon Mechanical Turk as well. However, the paper by Chakrabarty et al. (2022) does not detail whether any controls were applied or not for the selection of crowd workers. Nor were any details provided about the cohort of participants in terms of demographic data and the total number of participants recruited.

[2] ReproGen - https://reprogen.github.io/



"I ...I'll ask you once more. How much did you make by a fighter taking that dive?" The boy hesitated. "A thousand dollars." Lucky Luciano laughed. "That's **chicken feed**".

He was lucky he made so little

Lucky Luciano made much more money by doing other things such as fixing cars or selling drugs

She wanted to know if it could do more damage. "It's so hard to say," he answered. "But I honestly don't know how it could do more damage. Maybe it could deflect the attention to him, where it should be." "it's fine," she said, her monotone voice an echo of his sister. She was like **a high-definition TV**.

It was hard to tell what she was thinking

Her voice was so smooth, you could hear every word she said

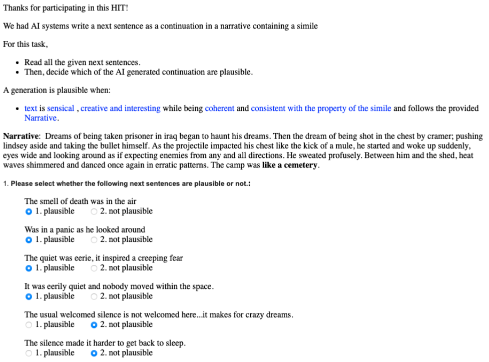
Figure 1: Diagram illustrating the judging of whether a given continuation is plausible or not with the top being an idiom and the bottom a simile. Taken from Chakrabarty et al. (2022).

For the reproduction experiment a total of 80 workers were recruited across both tasks (35 for idiom and 45 for simile). In agreement with the ReproHum organisers each worker was paid the UK living wage[3] of £10.90, in a US dollar equivalent amount, to give fair compensation for the workers time and effort across both tasks.

The experimental data and user interface was taken from the original published source code repository[4]. However several challenges were encountered in attempting to reuse the original experimental data and user interface:

- The CSV data used to prepare the idiom and simile tasks (HITs) on the MTurk platform were not present in the authors code repository.
- The interface for the idiom plausibility task was missing and not present in the code repository.
- The interface for the simile plausibility task, whilst present, was incomplete due to CSS code being commented out in the file. This left a visually inadequate interface as show in figure 2.

To re-create the CSV files needed for the plausibility idiom and simile tasks on MTurk the output JSON files from the original experiment were used. In particular, the narrative, the automatic and human continuation texts for each of the scenarios were extracted from these JSON files using a Python script.



Thanks for participating in this HIT!

We had AI systems write a next sentence as a continuation in a narrative containing a simile

For this task,

- Read all the given next sentences.
- Then, decide which of the AI generated continuation are plausible.

A generation is plausible when:

- text is sensical , creative and interesting while being coherent and consistent with the property of the simile and follows the provided Narrative.

Narrative: Dreams of being taken prisoner in iraq began to haunt his dreams. Then the dream of being shot in the chest by cramer; pushing lindsey aside and taking the bullet himself. As the projectile impacted his chest like the kick of a mule, he started and woke up suddenly, eyes wide and looking around as if expecting enemies from any and all directions. He sweated profusely. Between him and the shed, heat waves shimmered and danced once again in erratic patterns. The camp was **like a cemetery**.

1. **Please select whether the following next sentences are plausible or not.:**

The smell of death was in the air
● 1. plausible ○ 2. not plausible

Was in a panic as he looked around
● 1. plausible ○ 2. not plausible

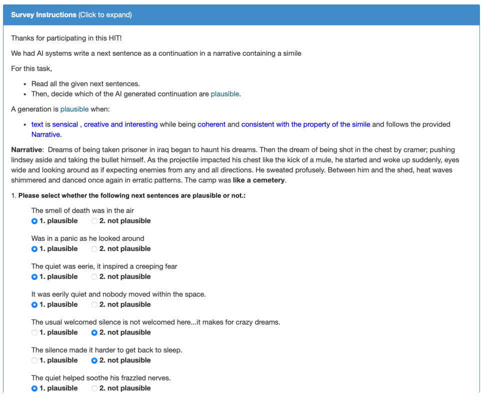
The quiet was eerie, it inspired a creeping fear
● 1. plausible ○ 2. not plausible

It was eerily quiet and nobody moved within the space.
● 1. plausible ○ 2. not plausible

The usual welcomed silence is not welcomed here...it makes for crazy dreams.
○ 1. plausible ● 2. not plausible

The silence made it harder to get back to sleep.
○ 1. plausible ● 2. not plausible

Figure 2: The simile plausibility interface with the missing CSS styling.



Survey Instructions (Click to expand)

Thanks for participating in this HIT!

We had AI systems write a next sentence as a continuation in a narrative containing a simile

For this task,

- Read all the given next sentences.
- Then, decide which of the AI generated continuation are plausible.

A generation is plausible when:

- text is sensical , creative and interesting while being coherent and consistent with the property of the simile and follows the provided Narrative.

Narrative: Dreams of being taken prisoner in iraq began to haunt his dreams. Then the dream of being shot in the chest by cramer; pushing lindsey aside and taking the bullet himself. As the projectile impacted his chest like the kick of a mule, he started and woke up suddenly, eyes wide and looking around as if expecting enemies from any and all directions. He sweated profusely. Between him and the shed, heat waves shimmered and danced once again in erratic patterns. The camp was **like a cemetery**.

1. **Please select whether the following next sentences are plausible or not.:**

The smell of death was in the air
● **1. plausible** ○ **2. not plausible**

Was in a panic as he looked around
● **1. plausible** ○ **2. not plausible**

The quiet was eerie, it inspired a creeping fear
● **1. plausible** ○ **2. not plausible**

It was eerily quiet and nobody moved within the space.
● **1. plausible** ○ **2. not plausible**

The usual welcomed silence is not welcomed here...it makes for crazy dreams.
○ **1. plausible** ● **2. not plausible**

The silence made it harder to get back to sleep.
○ **1. plausible** ● **2. not plausible**

The quiet helped soothe his frazzled nerves.
● **1. plausible** ○ **2. not plausible**

Figure 3: The simile plausibility interface with the restored CSS styling.

The missing CSS for the simile plausibility task was simply dealt with by re-instating the CSS by uncommenting the code in the interface HTML file resulting the interface as shown in figure 3. As for the missing interface file, after consultation with an organiser from the ReproHum project, a decision was made to copy the interface used from the simile task and make the relevant adaptions for the idiom task. In particular, this involved reducing the number of human text options from five to three

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

[4] Figurative Narrative Benchmark - https://github.com/tuhinjubcse/FigurativeNarrativeBenchmark

and using a randomly picked narrative text from the development set, and amending any mention of similes in the interface code to that of idioms.

Due to the limitations mentioned above, this reproduction experiment cannot be an exact replication of the experiment as conducted by Chakrabarty et al. (2022). Therefore the results presented later in this paper must take these limitations into account when considering any potential differences in the results obtained.

5 Results

Model	Original		Reproduction	
	Idiom	Simile	Idiom	Simile
GPT2-XL	56	60	58	64
+Context	68	68	83.33	48
+Literal	48	76	66.66	64
Human	80	88	80.55	84

Table 1: Percentage of model and human generated texts were majority rated as plausible by human evaluators. Original results are from (Chakrabarty et al., 2022).

Out of the 45 workers who participated on the simile plausibility evaluation only 5 workers had answered all 25 texts. In the idiom plausibility evaluation out of the 35 workers that had participated only 2 had completed all 25 texts, with the next highest participant completing 17 in total. Whilst this is lower than the original experiment, we believe this should not affect the results reported significantly as the analyses for idioms were constrained to 17 instead of 25 texts. Additionally, as shown later, a similar percentage of the idiom human texts were rated as plausible as the original study. For the analysis, the code was written independently from scratch as no analysis code is present within the authors code repository.

Table 1 shows the results from this analysis and results obtained from the reproduction study for each of the different text types that were rated plausible by a majority of human evaluators. We were able to get near exact or very close replication results for human and the baseline (GPT2-XL) generated texts. However, majority preference for the context and literal model texts are substantially different from the results reported by Chakrabarty et al. (2022).

When analysing inter-annotator agreement, the difference between the original study and the reproduction is a drop in the absolute Krippendorf's α score from 0.68 to 0.39. More granular analysis showed that the Krippendorf's α inter-annotator agreement was 0.3761 for the idiom task and 0.3971 for the simile task between the three respective annotators. It is possible that a difference in the type of annotators used in reproduction as compared the original study resulted in a difference in the inter-annotator results seen between two the studies.

When evaluating majority worker preference between the context model and the literal model for idioms and similes, we observed that for idioms preference was greater with the context model (83.33%) than the literal model (66.66%). For similes we were able to see a larger preference for the literal model (64%) over the context model (48%). This confirms the preferences that Chakrabarty et al. (2022) observed with human annotators in their original experiment.

Whilst we could not replicate the moderate inter-annotator agreement found in the original study nor the preference for the context model for idioms, we were able to successfully replicate the results for the percentage of idioms and similes considered plausible through majority worker voting for human and the baseline model generated texts. Additionally, we were able to replicate the preference for the context model for idioms and the literal model for similes. The fact that the results were either the same or very close to the original study shows in some aspects shows that some results were successfully replicated in this reproduction study.

6 Conclusion & Recommendations

In this paper we have conducted a partially successful reproduction of the results obtained in the absolute idiom and simile human evaluations. Whilst we were able to reconfirm the results for the judgement on whether human and baseline model generated texts idioms and similes are plausible, the same could not be said for the literal and context model texts. Additionally, inter-annotator agreement scores show that there is a significant difference between the results obtained as compared to the original study. One possible reason for this could be due to the difference in the cohorts of annotators recruited between the two studies. A similar challenge was found in the reproduction experiment by Mahamood (2021) where the difference in recruited participant cohorts was speculated as a possible probable cause for the inability to reproduce results from the original study. Nevertheless, it has been noted that recruiting MTurk crowd workers that have high inter-annotator agreement with each other can be challenging even with a structured process in place to filter out unsuitable workers (Zhang et al., 2023) and therefore in itself may not guarantee reproduction success.

Based on the experiences of this reproduction study there are several key recommendations to reduce uncertainty for reproduction attempts:

1. Give information on the type of participants in a given evaluation such as including demographic data.
2. State the inclusion and exclusion criteria for participants in an evaluation.
3. Provide the datasets, including any data preparation code, used to create crowd worker tasks and the respective task interfaces.
4. The analysis code used to compute the results from an evaluation should be included in the experiment's source code repository.

The first recommendation is very straightforward. Without the information on the type of participants that were used for the evaluation it is very likely that any reproduction attempt may not succeed as the differences between the two recruited groups might be too far significant to enable a comparable evaluation. Therefore, data, such as participant demographics, would enable any reproduction attempt to focus on recruiting the right participants for a given study. When combined with the second recommendation, this would help to give confidence to ensure that participants who do not qualify for the experiment are rightfully excluded. Once having recruited the right participants, it is important the exact same datasets and user interfaces are provided to ensure comparability with the original experiment and armed with the same analysis code to reduce any possibilities of discrepancies occurring. With these recommendations and the learnings from others in this area, it is hoped that future attempts at performing reproduction experiments will be more successful than at present.

Acknowledgments

Many thanks to *Evgeniya Pushenko* and *Srinivas Ramesh Kamath* of trivago for their time in reviewing this paper and for the improvements suggested.

References

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.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2020a. 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. 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, Simon Mille, and David M. Howcroft. 2020b. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. 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. Association for Computational Linguistics (ACL).

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.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

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.

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.

Saad Mahamood. 2021. Reproducing a comparison of hedged and non-hedged NLG texts. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 282–285, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. 2023. A needle in a haystack: An analysis of high-agreement workers on MTurk for summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14944–14982, Toronto, Canada. Association for Computational Linguistics.