ReproHum #0087-01: Human Evaluation Reproduction Report for *Generating Fact Checking Explanations*

Tyler Loakman[1], Chenghua Lin[2]
[1]Department of Computer Science, The University of Sheffield, UK
[2]Department of Computer Science, The University of Manchester, UK
tcloakman1@sheffield.ac.uk, chenghua.lin@manchester.ac.uk

Abstract

This paper presents a partial reproduction of *Generating Fact Checking Explanations* by Atanasova et al. (2020) as part of the ReproHum (Belz and Thomson, 2024) element of the ReproNLP shared task to reproduce the findings of NLP research regarding human evaluation. This shared task aims to investigate the extent to which NLP as a field is becoming more or less reproducible over time. Following the instructions provided by the task organisers and the original authors, we collect relative rankings of 3 fact-checking explanations (comprising a gold standard and the outputs of 2 models) for 40 inputs on the criteria of *Coverage*. The results of our reproduction and reanalysis of the original work's raw results lend support to the original findings, with similar patterns seen between the original work and our reproduction. Whilst we observe slight variation from the original results, our findings support the main conclusions drawn by the original authors pertaining to the efficacy of their proposed models.

Keywords: ReproNLP, Replication, Human Evaluation

1. Introduction

Recently, many works have investigated the role of human evaluation in assessing the quality of outputs in the field of Natural Language Processing (NLP) and Natural Language Generation (NLG) (Belz et al., 2023; Clark et al., 2021; van der Lee et al., 2019). Whilst human evaluation is often seen as the gold standard method of evaluation which takes into account the perceptions of real human end-users, there is much debate over the reproducibility of such evaluation (Belz et al., 2023; Howcroft et al., 2020). Automatic metrics, whilst scalable, frequently demonstrate poor concurrent validity, correlating poorly with human judgements (Liu et al., 2024; Zhao et al., 2023; Alva-Manchego et al., 2021; Reiter, 2018; Belz and Reiter, 2006). However, the performance of human evaluation has likewise been shown to have multiple flaws, including ill-defined evaluation criteria compounded by the absence of sufficient evaluator/annotator training to attenuate the subjectivity of the texts being rated from the subjective interpretation of the evaluation criteria itself. Furthermore, several works have discussed the presence of poorly selected human panels, including sufficient language proficiency and task understanding Schoch et al. (2020). This is further hindered by the choice of many works to obfuscate these shortcomings by neglecting to report any demographic information regarding participants, including for highly subjective language types such as humour (Loakman et al., 2023). Such discrepancies have resulted in widespread troubles in reproducing the results of different works in NLP (Thomson et al., 2024).

It is for reasons such as these that the ReproHum shared task aims to shine a spotlight on the level of reproducibility within the field of NLP through the mass reproduction of contemporary research through its many partner labs so that poor practices are identified and a record can be made of the progress of reproducibility over time, as researchers become increasingly aware of the best practices to follow in performing human evaluation in their works.

2. Background

As participants in the ReproHum project, we selected the paper *Generating Fact Checking Explanations* by Atanasova et al. (2020) as the focus of our reproduction, owing to interest in the topic of explanation generation, and previous experience of being part of evaluator panels for similar research. Through the automatic selection process, the ReproHum team identified the single experiment and criterion that we were to attempt to reproduce the results from, as introduced in §4.

Owing to our participation in the ReproHum project (Belz and Thomson, 2024), we were provided with the following materials: (i) a guide to the common approach to reproduction, (ii) the original paper and dataset required to perform a reproduction, and (iii) additional documents pertaining to clarifications and additional information provided by the original authors once contacted. During this process, the authors of this paper (and therefore the team performing the reproduction) did not con-

tact the authors of the original work directly at any stage.

In performing this reproduction, we adhered to the following criteria outlined in the documentation provided by the ReproHum organisers. All participants were paid minimally to the UK National Living Wage (12GBP per hour) as set by the ReproHum team for pair pay, in which we specifically paid 15GBP for this task and paid via Amazon Vouchers from our estimation that the task would take approximately 1.25hrs (which was confirmed by our evaluators following completion). Additionally, this work underwent ethical review and approval by the ethics review board of the primary author's institution (where all participants in this reproduction were also selected).

3. Original Study

In recent years with the widespread sharing of misinformation and the coining of "fake news", the need for accurate and reliable fact-checking systems has grown exponentially. While existing systems have demonstrated impressive performance, their "black box" nature often obscures the reasoning behind their predictions. This lack of transparency can hinder user trust and limit the adoption of these systems. Atanasova et al. (2020) identified an overall research focus on the veracity prediction task of news claims in existing research and a lack of work focusing on generating natural language explanations to justify these veracity predictions. They aimed to address the main drawback of a black-box system by generating explanations to support the assigned veracity labels. To do this, the authors leverage detailed fact-checking reports (termed "ruling comments") published alongside veracity labels by fact-checking organisations to produce explanations that resemble human-written justifications. This approach is further bolstered through a multi-task learning framework, where explanation generation is jointly optimised with a veracity prediction task for a DistilBERT (Sanh et al., 2020) based model. This joint training enables the system to identify regions in the ruling comments that are not only close to the gold standard explanation but also contribute to the overall fact-checking decision.

3.1. Evaluation

The authors evaluate their approach using both automatic and human evaluation methods. While automatic evaluation relies on the standard metric of ROUGE (Lin, 2004), human evaluation focuses on a range of different criteria listed below, alongside their original definitions:

- **Coverage** - The explanation contains important, salient information and does not miss any important points that contribute to the fact check.
- **Non-redundancy** - The summary does not contain any information that is redundant/repeated/not relevant to the claim and the fact check.
- **Non-contradiction** - The summary does not contain any pieces of information that are contradictory to the claim and the fact check.

Based on these criteria, evaluators are requested to rank different explanations based on their performance on each criterion (as well as providing an *Overall* ranking). The original results in Atanasova et al. (2020) demonstrate that the multi-task learning approach leads to improved performance for both veracity prediction and explanation generation. Notably, the generated explanations achieve better coverage and overall quality compared to explanations trained solely to mimic human justifications. This suggests that the joint training framework allows the system to capture the knowledge required for accurate fact-checking, leading to more informative and relevant explanations. In our reproduction, we focus solely on the underlined criterion of *Coverage*.

4. Reproduction Setting

Task Setting As directed by the ReproHum team, we performed our reproduction on a single element of the original work by Atanasova et al. (2020) regarding evaluating outputs on the aforementioned criteria of *Coverage*. We presented the same instructions to participants as presented by Atanasova et al. (2020) with minor changes, as presented in Figure 1. These changes exclusively involve the removal of information regarding other evaluation criteria used in the original study outside of *Coverage*, including *Non-redundancy*, *Non-contradiction*, and a holistic *Overall* rating. We additionally remove all mention of the separate Task 2 which is not the subject of this reproduction. As with the original study, we performed our reproduction experiment by having participants place their relative preference rankings of 3 systems (i.e., a gold standard and two models) in a spreadsheet facilitated via Google Sheets. Within this, 3 columns follow the 3 explanations (from the 3 different models) to place rankings (where the n-th column contains the ranking for the n-th justification), as outlined in Figure 1. In line with the recommended approach to performing reproductions presented by the ReproHum team, we additionally incorporate data validation techniques in the form of drop-down

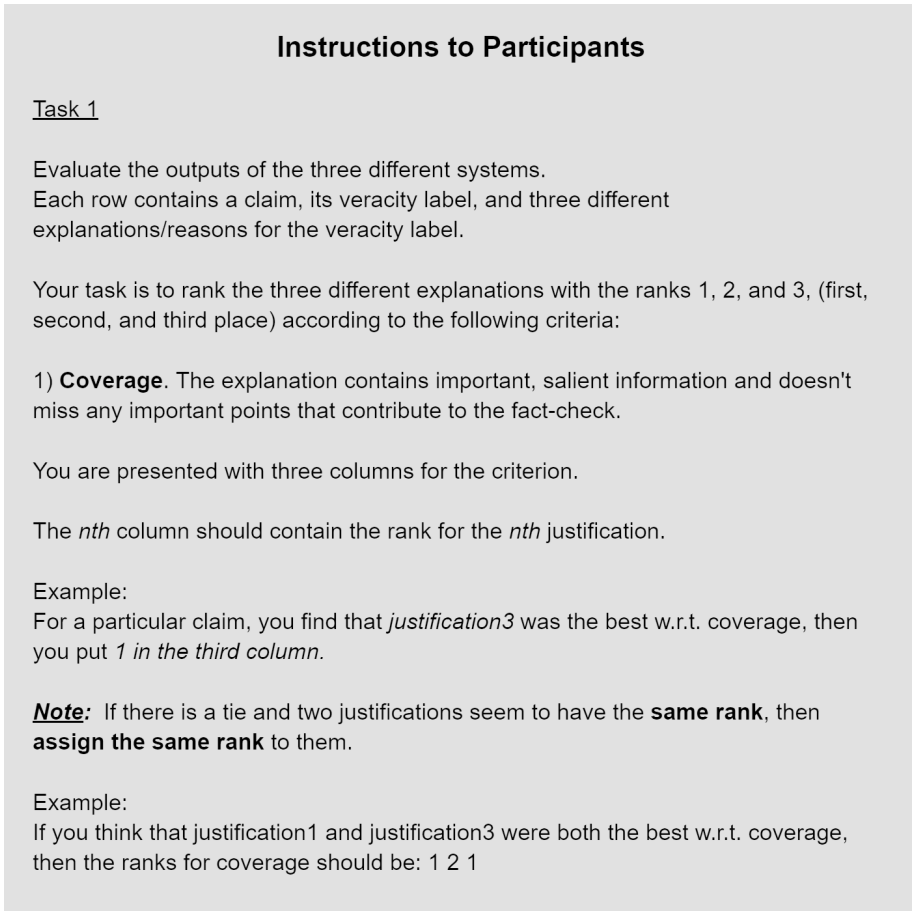

Figure 1: Modified instructions from Atanasova et al. (2020) presented to participants within the reproduction. we made minor modifications to the original instructions presented to participants in order to remove information related to tasks and criteria that were not to be assessed in this reproduction.

boxes containing rankings of 1-3. This ensured that participants only entered valid options in the ranking task. We present model outputs to participants in the same shuffled order presented in the original paper to also avoid order effects and bias towards particular columns. In total, each participant annotated 120 items, consisting of the outputs of 3 systems (including the human gold standard) for 40 inputs. We also make available a HEDS datasheet (Shimorina and Belz, 2022) detailing the process of our reproduction study.[1]

Evaluator Demographics In the original work by Atanasova et al. (2020) we have limited demographic details regarding the participants. However, we are aware that they are colleagues of the authors and have experience in fact-checking annotation tasks, whilst not exclusively being native speakers of the target language. In our replication, we use 3 Ph.D. students in Natural Language Processing, all of which have experience in fact-checking and related tasks (e.g., misinformation/rumour detection). All participants in our reproduction also have a professional working level of English fluency.

5. Results

We present the results of the original study and our reproduction in Table 1. Due to minor discrepancies in the specific evaluated materials (owing to some evaluators in the original work assessing approximately 80 items, and others assessing only 39, with some omissions), we additionally report what we term a "recreation", where we reanalyse the original paper's raw data to facilitate a direct comparison against only the same 40 inputs as presented to our evaluators. In the original work by Atanasova et al. (2020), the criterion of *Coverage* is shown to have low inter-annotator agreement as calculated via Krippendorff's Alpha (Krippendorff, 2019), reporting $\alpha = 0.26$ across their 3 evaluators. In our reproduction, we find slightly better agreement among our participants, with $\alpha = 0.35$ when specifically accounting for an ordinal level of measurement, whilst we find agreement across the 40

[1]Available at https://github.com/nlp-heds/repronlp2024.

Original			
Annotators	**Gold**	**Explain-Extr**	**Explain-MT**
All	**1.48**	1.89	1.68
1st	**1.50**	2.08	1.87
2nd	**1.74**	2.16	1.84
3rd	**1.21**	1.42	1.34
CV*	9.00%	8.10%	5.76%
Recreation			
Annotators	**Gold**	**Explain-Extr**	**Explain-MT**
All	**1.52**	1.87	1.66
1st	**1.55**	2.05	1.85
2nd	1.82	2.15	**1.77**
3rd	**1.18**	1.41	1.36
CV*	6.35%	9.16%	6.96%
Reproduction			
Annotators	**Gold**	**Explain-Extr**	**Explain-MT**
All	**1.62**	2.05	1.78
1st	**1.60**	2.30	2.03
2nd	1.60	1.86	**1.55**
3rd	**1.65**	1.98	1.75

Table 1: Comparison between Atanasova et al. (2020) and our reproduction on the criterion of "Coverage". Values present the Mean Average Ranks (MAR) of the explanations. The explanations come from the gold justification (**Gold**), the generated explanation (**Explain-Extr**), and the explanation learned jointly (**Explain-MT**) with the veracity prediction model. A lower MAR indicates a better average ranking. For each row, the best results are in **bold**, and the best automatically generated explanations are underlined. "Annotators" refers to each individual rater, whilst "All" is the mean across all annotators. *CV** refers to the Coefficient of Variation for the mean ratings of the 3 systems compared to our reproduction results following the implementation by Belz (2022). *Original* refers to the results presented in the original paper by Atanasova et al. (2020), whilst *Recreation* refers to the results we gain by reanalysing the original study's data exclusively for the same sample that our evaluators were presented. Finally, *Reproduction* refers to the results of our reproduction study using our new evaluators. The ordering of annotators across *Recreation* and *Original* should be considered arbitrary, as we cannot guarantee each line corresponds to the same annotator as the original.

evaluated inputs in the original data to be very similar to what was reported for the particular subset used by the authors in the original work ($\alpha = .27$)

In terms of overall patterns seen in the data, the results of our reproduction can be seen to differ slightly from those of the original in terms of overall rankings. Firstly, in the original study, the golden human-authored explanations were preferred by all participants, whilst this is not seen to be the case in our reproduction or in our reanalysis of a specific subset of the original paper's raw data (i.e., *recreation*) Instead, we find only 2 of our 3 participants to rank the golden explanations in their expected 1st place. However, in terms of the automatically generated explanations we observe the *Explain-MT* model (where the explanation is learnt jointly with the veracity prediction model) to outperform *Explain-Extr* (where the auxiliary veracity prediction model is learnt separately), mirroring the results presented in the original work.

Furthermore, when aggregating the results of all 3 evaluators in our reproduction, we can see that the overall rankings assigned to each output are higher (i.e., worse) than the findings of Atanasova et al. (2020). However, whilst our raw figures differ from the original findings (owing to the relatively subjective task criteria and small evaluator panel sizes), our findings reflect the same overall patterns as the original work, with the human-authored golden explanations *Gold* outperforming the authors' proposed models in the majority of cases, whilst the more complex *Explain-MT* model, which is trained alongside a veracity prediction task, outperforms the *Explain-Extr* model that learns to generate explanations in isolation.

To compare against the original study's findings, we calculate correlations between our results and those provided by the original paper's authors using Spearman's ρ and Pearson's r. Due to the original work's raw data having results for more than 40

trials, and with some missing values, we assess only the same 40 trials as presented to our participants (equivalent to the *Recreation* in Table 1) and calculate the mean rank given to each output by the evaluators (which is robust to cases where not all evaluators in the original work assessed a given output). The results show a strong correlation between the results of our reproduction and the original study ($\rho = .524$ and $r = .541$, which are both significant at $\alpha = .01$), demonstrating that we were able to reproduce the general evaluator preferences observed in the original experiment.

6. Conclusion

In this paper, we have presented our reproduction findings for an element of human evaluation presented in Atanasova et al. (2020) regarding the criteria of *Coverage* to compare gold standard fact-checking explanations with 2 proposed models. In terms of overall comparison with the original work, we find a higher level of rating agreement among our evaluator panel than demonstrated in the original work but also observe a slightly different overall pattern than presented by the original authors, with one of the proposed models ranking higher than the gold standard human-authored explanation from 1 of our 3 participants. We do, however, observe the same pattern when reanalysing the raw data from the original study, focussing exclusively on the same subset of examples presented to our evaluators in the reproduction. Additionally, our reproduction lends credence to the results presented by Atanasova et al. (2020) regarding the model trained to generate explanations alongside a veracity prediction model (Explain-MT) outperforming the model that is trained to generate explanations in isolation (Explain-Extr) in terms of human rankings. It is important to note, however, that the result of our reproduction covers only one of the multiple human evaluation criteria on which the raters were asked to assess the generations in the original work, and this pattern may not necessarily be present across all different criteria.

Overall, we reiterate the importance of performing reproduction studies such as this in order to assess the trend of reproducibility within the field of NLP. Within this paper, we have successfully reproduced the findings of the original work with some minor variability (likely owing to the small size of the evaluation panels in the original work, and consequently our reproduction). This is particularly salient for the topic of generating fact-checking explanations that Atanasova et al. (2020) tackle, as this constitutes a high-impact application of NLP with an increased need for reliable and robust models and evaluation procedures in order to avoid the effects of misinformation.

7. Acknowledgements

We would like to thank the organisers of the ReproHum/ReproNLP shared task for their efforts in bringing this large-scale reproduction effort to light and for their continued assistance throughout the process of performing our reproduction experiments. We would also like to thank the authors of *Generating Fact Checking Explanations* (Atanasova et al., 2020) for their transparency in providing the necessary resources and materials to aid in our reproduction of their work.

Tyler Loakman is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications, funded by UK Research and Innovation [grant number EP/S023062/1]

8. Bibliographical References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

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

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.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. 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.

Klaus Krippendorff. 2019. Content analysis: An introduction to its methodology.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2024. Llms as narcissistic evaluators: When ego inflates evaluation scores.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.

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

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

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Kun Zhao, Bohao Yang, Chenghua Lin, Wenge Rong, Aline Villavicencio, and Xiaohui Cui. 2023. Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–574.