ReproHum #0087-01: A Reproduction Study of the Human Evaluation of the Coverage of Fact Checking Explanations

Mingqi Gao, Jie Ruan, Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
{gaomingqi, wanxiaojun}@pku.edu.cn
ruanjie@stu.pku.edu.cn

Abstract

We present a reproduction study of the human evaluation of the coverage of fact checking explanations conducted by Atanasova et al. (2020), as a team in Track B of ReproNLP 2024. The setup of our reproduction study is almost the same as the original study, with some necessary modifications to the evaluation guideline and annotation interface. Our reproduction achieves a lower IAA of 0.20 compared to the original study's 0.27. Additionally, our reproduction results on the ranks of three types of explanations are drastically different from the original experiment, rendering that one important conclusion in the original paper cannot be confirmed at all. The case study illustrates that the annotators in the reproduction study may understand the quality criterion differently from the annotators in the original study.

Keywords: reproduction study, human evaluation, fact checking explanations

1. Introduction

These years have witnessed the concern about reproducibility issues in the field of NLP, especially human evaluation (Belz et al., 2023). In this paper, we present a reproduction study of human evaluation of the coverage of fact checking explanations (Atanasova et al., 2020), as a team in the Track B of ReproNLP Shared Task 2024 (Belz and Thomson, 2024).

The original study (Atanasova et al., 2020) formalizes fact checking as follows: Given a claim and some ruling comments, the model is required to predict the veracity label of the claim and also generation explanations. In the original experiments, human evaluation was performed to compare the quality of gold explanations and the explanations generated by two proposed models. The explanations were ranked by human annotators according to four quality criteria separately: Coverage, Non-Redundancy, Non-Contradiction, and Overall. After a discussion with the organizers of ReproNLP, we are asked to conduct a reproduction study only for Coverage.

2. Experimental Design

2.1. Original Experiment

LIAR-PLUS (Alhindi et al., 2018), a fact checking dataset based on PolitiFact [1], was used in the original study. Each instance of the dataset contains a claim, some ruling comments, a veracity label, an automatically extracted justification as the gold explanation, and other metadata (e.g. speaker). There are six veracity labels: pants-fire, false, mostly false, half-true, mostly-true, and true.

The gold explanations in the dataset are abbreviated as **Just** in the original study. Besides, two explanation generation models are proposed: **Explain-MT** was trained jointly with veracity label prediction and **Explain-Extr** was trained separately.

Selection of evaluation instances. According to the original paper, 40 instances were randomly selected from the test set and three veracity explanations were collected for each of them. Each instance for human evaluation includes an instance ID, a claim, a veracity label, and three explanations. The ruling comments are excluded. Additionally, it is worth mentioning that after examining the original annotation interface (the Excel file), we find there are 80 instances included. Nevertheless, according to the raw annotation in the original experiment, only the first half was annotated by all three annotators.

Participating annotators and compensation. It is reported in the original paper that three annotators were involved but other information is not mentioned. According to the materials provided by the organizers of ReproNLP, none of the annotators were English native speakers. They were all colleagues of the authors and had previous experience with fact checking tasks. There is no information on whether and how much they were paid.

Quality criterion. The definition of the coverage of the explanation is as follows:

Coverage. The explanation contains important, salient information and doesn't miss any important

[1] https://www.politifact.com/

id	claim	LABEL	justification 1	justification 2	justification 3	Coverage			Non-redundancy			Non-contradictory			Overall		
2568.json	impleme	FALSE	sure that an	the Grand	lower the	1	2	2	1	2	2	1	2	2	1	2	2
11923.json	will work	half-true	checking if	she will "work	Medium post,												
11025.json	thousan	FALSE	honest but	her state of	would have												
10085.json	on	FALSE	Mount	Mount	independent,												
9622.json	women	TRUE	ton of	ton of women	said that												
7834.json	in the	TRUE	matter, and	ng that the	the right to												
2205.json	106,000	TRUE	d health and	that we had	that we had												
8606.json	ans have	half-true	the health	the health	said,												
575.json	McCain	barely-true	'intervening'	Airbus get	say two												

Figure 1: Annotation interface used in the original experiment. There are 80 instances in total and only the first ten are shown.

id	claim	LABEL	justification 1	justification 2	justification 3	Coverage		
2568.json	impleme	FALSE	sure that an	Grand Canyon	lower the	1	2	2
11923.json	will work	half-true	checking if	she will "work	Medium post,			
11025.json	thousand	FALSE	honest but	her state of	would have			
10085.json	on Mount	FALSE	Mount	Mount	independent,			
9622.json	women	TRUE	ton of women	ton of women	said that			
7834.json	in the	TRUE	matter, and	g that the right	the right to			
2205.json	106,000	TRUE	health and	we had over	we had over			
8606.json	ans have	half-true	the health	the health	"Republicans			
575.json	McCain	barely-true	'intervening' is	Airbus get the	say two			

Figure 2: Annotation interface used in our reproduction experiment. There are 40 instances in total and only the first ten are shown.

points that contribute to the fact-check.

Evaluation methods. Given three different explanations (**Just**, **Explain-Extr**, and **Explain-MT**), the annotators were asked to rank 1,2,3 according to the criterion. It is noted in the evaluation guideline that if there is a tie and two explanations seem to have the same rank, the annotation should assign the same rank to them.

Annotation interface. The annotation was conducted through an Excel file, a screenshot of which is shown in Figure 1. In each row, the three explanations were randomized in terms of where they were placed to ensure fairness. Annotators were asked to record their ranks of the three explanations in the same row.

Annotation procedure. According to the information provided by the organizers of ReproNLP, there is no training process. Three participants were asked to read the evaluation guideline and then annotate the selected 40 instances separately.

Inter-annotator agreement (IAA). Krippendorff's α (Hayes and Krippendorff, 2007) was used to measure the IAA.

Presentation of results. For each type of explanation, the mean average ranks (MAR) by each annotator were presented. The average MAR of the three annotators was taken as the final result.

2.2. Reproduction Experiment

We were provided with an Excel file that included all the evaluation instances and an evaluation guideline. Both of them are exactly the same as the original experiment, which makes the setup of our reproduction experiment almost identical to the original experiment. The main differences from the original experiment are described below. For more details, please refer to the Human Evaluation Sheet (HEDS) (Shimorina and Belz, 2022) in supplementary materials [2].

Modifications to the evaluation guideline and the annotation interface. In the original study, in addition to Coverage, the annotators needed to assess the explanations against each of the three other quality criteria: Non-Redundancy, Non-Contradiction, and Overall. Additionally, there is another human evaluation task in the original study:

[2]They are also available at https://github.com/nlp-heds/repronlp2024.

	Original	Reproduction	Confirmation
1	The gold explanation ranks the best in Coverage.	The gold explanation ranks the worst in Coverage.	Not confirmed.
2	*Explain-MT* ranks better than *Explain-Extr* in Coverage.	*Explain-MT* ranks better than *Explain-Extr* in Coverage.	Confirmed.

Table 1: The conclusions from the original paper and the conclusions according to our reproduction results. The confirmation column shows whether the conclusion in the original study is confirmed or not.

	Just	Explain-Extr	Explain-MT
original (calculated by us) vs. original (from the paper)	0.68	0.53	1.18
reproduction vs. original (from the paper)	38.14	2.09	3.63
reproduction vs. original (calculated by us)	38.79	2.62	4.80

Table 2: CV*s among different experiment results. The smaller the CV*, the closer the results.

original (calculated by us) vs. original (from the paper)	1.00
reproduction vs. original (from the paper)	-0.50
reproduction vs. original (calculated by us)	-0.50

Table 3: System-level Spearman's ρ among different experiment results.

	nominal	ordinal	interval	ratio
Original (calculated by us)	0.16	0.27	0.27	0.26
Reproduction	0.12	0.20	0.20	0.18

Table 4: Krippendorff's α. Different columns denote the annotations are viewed as nominal, ordinal, interval, or ratio data. In general, ranks are considered ordinal data.

providing the veracity label based on the explanations. These are reflected in the original evaluation guideline and the Excel file. We removed the content about other quality criteria and tasks from the evaluation guideline and the Excel file because we only reproduced the coverage evaluation of the explanations. The original evaluation guideline and the modified guideline are both included in the supplementary materials. The modified Excel sheet is shown in Figure 2. Furthermore, we only include the first 40 instances in our Excel file.

Participating annotators and compensation. Following the discussion with the organizers of ReproNLP, we recruited three PhD students who were proficient in English and paid them 12.24 EUR per hour.

3. Results

In addition to the evaluation guideline and the Excel file for annotation, we were also provided with the raw annotation of each annotator in the original experiment, which enabled us to calculate the IAA and the MAR based on the raw annotation. It is interesting to see whether it is different from what is reported in the paper. Therefore, we present

Annotators	Just	Explain-Extr	Explain-MT
Original (from the paper)			
All	**1.48**	1.89	1.68
Annotator #1	**1.50**	2.08	1.87
Annotator #2	**1.74**	2.16	1.84
Annotator #3	**1.21**	1.42	1.34
Original (calculated by us)			
All	**1.47**	1.88	1.70
Annotator #1	**1.48**	2.08	1.90
Annotator #2	**1.72**	2.15	1.87
Annotator #3	**1.21**	1.41	1.33
Reproduction			
All	2.18	1.93	**1.62**
Annotator #1	2.05	2.00	**1.65**
Annotator #2	2.23	1.95	**1.62**
Annotator #3	2.25	1.85	**1.60**

Table 5: Mean Avegae Ranks (MAR) of different explanations, the lower the better. The best result of each row is in **bold** and the best generated explanation is in lightgrey. The first four rows are directly taken from the original paper. Note that there is no correspondence between the annotators of the original experiment and our reproduction experiment.

three types of experiment results in this section: the original results taken from the original paper, the original results calculated by us with the raw annotation, and our reproduction results. The former two share common human annotation, and the latter two share a common calculation process.

3.1. Inter-annotator Agreement

The Krippendorff's α reported in the original paper is 0.26. It is close to the results (0.27) calculated by us with the raw annotation in the original experiment, as shown in Table 4. Based on our calculation, the IAA of our reproduction experiment is

Claim	Warren Buffett called (Jeff) Greenes scheme financial weapons of mass destruction.
Label	FALSE
Just	Still, there's no doubt Greene was heavily involved in buying credit default swaps, which are a type of derivative. And Buffett was exceptionally critical of derivatives. The problem with Meek's ad is that it gives the impression that Buffett criticized Greene personally, which he didn't do. By saying Buffett called Greene's scheme a bad idea, Meek personalizes a comment that had nothing to do with Greene.
Just MAR	Original: 1.00 (1, 1, 1), Reproduction: 3.00 (3, 3, 3)
Explain-Extr	Warren Buffett called Greene's scheme 'financial weapons of mass destruction.' We looked at Meek's statement that Greene "is the king of the undercover Credit Default Swap that brought about the destruction of our economy that we have right now." 317 on the Forbes list) would have made it onto Buffett's radar. We should note here that journalists love Buffett's phrase "financial weapons of mass destruction."
Explain-Extr MAR	Original: 2.67 (2, 3, 3), Reproduction: 1.33 (1, 1, 2)

Table 6: An example that shows the different annotation results between the original experiment and our reproduction experiment. The rows of MAR list how the three annotators rank the explanation and the mean average ranks.

lower than the original experiment.

The original paper considers a low IAA of 0.26 may be caused by the high subjectivity of ranking and the difficulty of this task. We believe that the inadequate evaluation guideline may also contribute to the low IAA. First, there is no example for each quality criterion. Second, the six veracity labels (pants-fire, barely-true, half-true, mostly-true, false, and true) lack clear definitions, which makes the evaluation of explanations harder.

3.2. Side-by-side Comparisons

Table 5 shows that there are minor differences in MAR between the results taken from the original paper and the results calculated by us with the raw annotation in the original experiment. However, **our reproduction results are dramatically different from the original experiments.** As shown in Table 1, a conclusion that the gold explanation ranks the best for Coverage is not confirmed at all, and our reproduction experiment yields the opposite conclusion. Despite this inconsistency, another conclusion is confirmed by our reproduction experiment.

We also present CV*, a metric proposed by Belz et al. (2022) to quantify reproducibility (in Table 2) and Spearman's ρ (in Table 3) among different experiment results, also demonstrating the small differences between the original results calculated by us and from the paper but sharp inconsistency between our reproduction experiment and the original experiment.

3.3. Discussion

Case study. The big difference in the ranks of the gold explanations (Just) encourages us to conduct a case study. After examining some instances that differ from the original annotations, we conclude that the annotators in the reproduction study may understand the quality criterion differently from the annotators in the original study. The annotators in the original study pay more attention to whether the veracity label can be inferred from the explanation, while the annotators in the reproduction study focus more on whether the information in the claim is covered by the explanation. Table 6 shows an example. The annotators' understanding in the original study may be more reasonable but the ambiguity in the definition of the quality criterion is also the cause of this phenomenon.

Explanation of the version correction. In the first version, we found that Krippendorff's α (0.12) calculated by us with the raw annotation in the original experiment is quite different from what is reported in the original paper (0.26). After a discussion with the partner lab, we realized this is because **the instance ID column of the annotator #3's raw data in the original study is problematic**. We used the wrong instance IDs to align the raw data of the three annotators and derived different results. We have updated all the results that could be affected by this issue.

4. Conclusion

In this paper, we present a reproduction study of the human evaluation of the coverage of fact checking

explanations under the guidance of the organizers of ReproNLP. Our conclusions are as follows:

- Our reproduction achieves a lower Krippendorff's α of 0.20 than the original experiment (0.27) based on our calculation, though both of them are not satisfactory.
- The results of our reproduction experiment are drastically different from the original experiment, rendering that one important conclusion in the original paper cannot be confirmed at all.
- There are minor differences between the results calculated by us with the raw annotation in the original study and the results reported in the original paper.

5. Bibliographical References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

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.

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. The 2024 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

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

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

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.