ReproHum #0043-4: Evaluating Summarization Models: Investigating the Impact of Education and Language Proficiency on Reproducibility

Mateusz Lango, Patrícia Schmidtová, Simone Balloccu, Ondřej Dušek
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{lango,schmidtova,balloccu,odusek}@ufal.mff.cuni.cz

Abstract

In this paper, we describe several reproductions of a human evaluation experiment measuring the quality of automatic dialogue summarization (Feng et al., 2021). We investigate the impact of the annotators' highest level of education, field of study, and native language on the evaluation of the informativeness of the summary. We find that the evaluation is relatively consistent regardless of these factors, but the biggest impact seems to be a prior specific background in natural language processing (as opposed to, e.g. a background in computer science). We also find that the experiment setup (asking for single vs. multiple criteria) may have an impact on the results.

Keywords: human evaluation, reproduction, reproducibility, dialogue summarization, summarization

1. Introduction

Human evaluation is generally considered to be the gold standard for Natural Language Processing (NLP) systems assessment. However, many factors can affect its reliability. Subjectivity in human ratings can make experiments impossible to reproduce (Belz et al., 2021); the adopted definition for the evaluated criteria can confuse the annotators (Hosking et al., 2024), and external factors (e.g. fluency) can influence them (Wu et al., 2023). As researchers, we often do not realize the flaws in our own evaluation schemes (Thomson et al., 2024), but they can be found when someone else tries to reproduce such evaluation. Therefore, efforts such as the ReproHum project (Belz et al., 2023a; Belz and Thomson, 2023) also help us design better and more robust human evaluation practices.

In this paper, we describe our attempt at reproducing the human evaluation experiment on dialogue summarization originally presented by Feng et al. (2021) (see Section 2). We specifically target reproduction on one of the datasets and focus mainly on the informativeness criterion. We set up an initial experiment with a setting as close as possible to the original study, including hiring the participants. We then run three variants of the reproduction, inspecting the effect of annotators native language and general background (including knowledge of NLP), using participants hired over the Prolific crowdsourcing platform (see Section 3).

Our reproductions were able to confirm some of the original paper's high-level conclusions from the human evaluation, but also showed some substantial differences in relative rankings among Feng et al. (2021)'s own systems as well as in absolute ratings of all evaluated summarizers (see Sections 4 and 5). The differences between individual reproductions indicate that while participants' native language and general background are not very relevant, a specific background in natural language processing (NLP) can have an impact. In addition, the particular setup of the evaluation (i.e., checking for single or multiple criteria) seems to affect the results (see Sections 6 and 7).

2. Original Experiment

The original paper (Feng et al., 2021) proposes a method on how to leverage DialoGPT (Zhang et al., 2020b) as a dialogue annotator to assist in the task of dialogue summarization. The annotations are added as a pre-processing step prior to the summarization.

The authors test their methods on two datasets: SAMSum (Gliwa et al., 2019) and AMI (Carletta et al., 2006). The performance is evaluated using a combination of automatic metrics – ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020a) – as well as human evaluation. The authors report that their additional DialoGPT-derived annotations are capable of improving the performance of a pre-trained summarizer – BART (Lewis et al., 2020) and a non-pre-trained summarizer – Pointer-Generator Networks (PGN, See et al., 2017) on both datasets. They also report a new state-of-the-art performance on the SAMSum dataset.

While the paper includes results with multiple external baselines and BART pre-trained model extensions, these are either only used with automatic metrics, or only on the SAMSum dataset. For human evaluation on the AMI dataset, which is rel-

Proceedings of the HumEval @LREC-COLING 2024, pages 229–237
Turin, Italy. May 21st, 2024. ©2024 European Language Resources Association (ELRA)

evant for our reproduction, the compared systems were:

- Hierarchical Meeting summarization Network (HMNet, Zhu et al., 2020) – a variant of the encoder-decoder transformer model, specially adapted for modelling dialogues. The network was pre-trained on news summarization data. The results of this model serve as a strong external baseline.
- Vanilla Pointer-Generator Network (PGN, See et al., 2017) is an LSTM-based model that combines standard encoder-decoder architecture with pointer network. No pretraining was applied. This model was used by the authors as a baseline summarizer, which was extended with different annotations proposed by the authors.
- PGN with keyword extraction annotation (D_{KE}) – the input to the PGN summarizer is extended with a list of keyword words extracted by analysing the outputs of DialoGPT.
- PGN with redundancy detection annotations (D_{RD}) – a special tag is added in front of each utterance in the dialogue that has been detected as redundant.
- PGN with topic segmentation annotations (D_{TS}) – a special tag is prepended to each utterance that starts a new topic in the dialogue as detected by DialoGPT.
- PGN with all the above annotations (D_{ALL})– the input to PGN is enhanced with all the additional annotations described above.

Additionally, one dialogue summary written by a human was evaluated for comparison.

Human evaluation is performed on summaries generated for 10 randomly selected dialogues. Four annotators are asked to rate informativeness, conciseness and coverage on a 5-point Likert scale, as well as provide a binary good/bad indication for each summary. More details on the experiment are discussed in Section 3, where we also describe key differences of our reproduction.

3. Reproduction Studies

We performed four reproductions of the experiment described above: one according to the ReproHum project guidelines (dubbed ReproHum) and three additional ones (dubbed Repro #1 through #3) to investigate different factors influencing the results of human evaluation. We have tried to follow the original experimental setup as closely as possible, but there are still several differences between the original experiment and our reproductions. We first describe the conditions for the ReproHum study, then detail how the additional studies differ from it.

3.1. ReproHum reproduction

The ReproHum reproduction used the following setup as a result of the original study's setup and ReproHum guidelines (Belz and Thomson, 2024):

Datasets The original experiment was performed on both SAMSum (Gliwa et al., 2019), and AMI (Carletta et al., 2006), datasets, but the reproduction was limited to the latter dataset only to limit cost. same 10 dialogues from AMI were used, presented in the same order.

Annotation interface The annotation interface was slightly different. In the original study, the authors used a simple text file to collect annotations – more specifically, they used a custom script that printed the dialogues on the console and then prompted the user to rate different summaries. We performed the annotations in a Google form, following ReproHum guidelines.

Evaluated quality factors The original annotations include three quality factors: informativeness, conciseness, coverage, and a final binary rating of whether the presented summary is good. Interestingly, the results of the final binary evaluation were not reported in the original study. In our reproduction, we limited the study to the evaluation of informativeness only, following the decision of the ReproHum team.

Annotators All the original annotators were Chinese PhD students with a background in NLP, specifically in text generation or summarization tasks. Their level of English was assessed by a Chinese state examination of English proficiency: College English Test (CET-6).

In our ReproHum reproduction, we aimed at getting a close demographic, with main focus on hiring PhD students. Our annotators were thus all PhD students and non-native speakers of English, hired on a contract basis. However, they had no background in computer science, their native languages did not include Chinese, and their fluency level was self-assessed.

Remuneration In the original study, the human evaluators were paid 10 USD each. According to the ReproHum fair pay policy, the reproduction wage was set at 14.3 USD per hour. The time needed to perform the annotation was estimated to be 8 hours, which resulted in a total wage of approx. 115 USD per annotator.[1]

Annotation guidelines The original study used annotation guidelines in Chinese. As the annotators in our reproduction did not speak Chinese, we translated the annotation guidelines into English. In addition, since our reproduction concerns only one quality factor and one dataset, the guideline was edited to remove mentions of other quality factors and the SAMSum dataset. The final annotation guideline is as follows:

> *Hi everyone, thanks for helping to do the human evaluation, there is one dataset, AMI, long conversation, 10 data items in total. Please mark each based on the indicator: Informativeness, ranging from 1 to 5, 1 being the worst 5 being the best. Informativeness measures whether the abstract contains the key information from the original conversation. Everyone's document is the same, a total of 4 people will evaluate the same data, and we will then calculate the kappa value to measure the consistency.*

3.2. Additional reproductions

We conducted three additional reproduction experiments to investigate the influence of annotators (a) having a background in computer science, (b) having English or Chinese as their first language, (c) answering all four questions as in the original experiment, instead of just one as in the ReproHum reproduction. We followed the same approach as in the ReproHum study, except for annotator demographics and the set of questions (Repro #3 only). We used the Prolific platform as an intermediary to easily find annotators with the necessary background.[2]

The specifics of the additional studies are as follows:

- **Repro #1** was conducted by annotators with a background in computer science (at least a bachelor's degree) and Chinese as their native language.
- **Repro #2** was performed by annotators with a background in computer science (at least having completed a bachelor's degree) and English as their native language.
- **Repro #3** was done by annotators with a background in computer science (at least a bachelor's degree) who were native Chinese speakers. The annotators were responding to all the questions from the original study.[3]

Contrary to the original study, our annotators did not have specific background in NLP and were not PhD students. This difference is given by limited annotator availability on the Prolific platform.

As the workload for an annotator was estimated at 8 hours, we decided to divide the study into 10 parts, corresponding to summaries of 10 evaluated dialogues. Each Prolific annotator was required to complete all parts of the study within a two-week period. Each reproduction was carried out with 4 annotators and the same remuneration as for the ReproHum reproduction.

The task of evaluating long dialogue summaries is not ideally designed for platforms such as Prolific. It relies on reading a long text[4] and then answering several questions on a 1-5 scale (or giving a binary response). There are no attention checks and it is rather difficult to design such. For instance, asking questions about dialogue content could inadvertently suggest to annotators that these parts of the dialogue were important and should be included in the summary. Therefore, we used the time spent on the task as a weak indicator of the annotator's careful reading and analysis of the dialogue content. According to Brysbaert (2019), the average adult has a reading speed of 175 to 300 words per minute (wpm), so annotators who completed the annotation of the first dialogue in a time corresponding to a theoretical reading speed of more than 400 wpm were rejected.

4. Main Results

The informativeness values obtained in our reproductions together with the results from the original study (Feng et al., 2021) are presented in Table 1. A rank analysis of these results (the higher, the better) is shown in Table 2.

Absolute score differences All the reproductions are very consistent with the original study in rating the informativeness of the human-written summaries highly, at a very similar level. On the contrary, in our reproduction all automatically generated summaries were rated substantially lower than in the original study. As this is consistent regardless

[1] The time estimation was done by a trial annotation of two summaries ran by the reproduction authors. The payments were handled in CZK, we provide conversions based on the exchange course as of March 2023 (1 USD = 23.4 CZK).

[2] https://app.prolific.com/

[3] Due to an error in one of the Google Forms, the question about the overall binary quality evaluation was omitted for one dialogue summary.

[4] The joint text of 10 dialogues and the corresponding summaries has almost 65,000 words, which corresponds to 159 A4 pages in 11pt Courier New font.

		Reproductions			
	Original	ReproHum	Repro #1	Repro #2	Repro #3
Evaluated factors	All	Inform.	Inform.	Inform.	All
Educational level	PhD Student	PhD Student	≥Bachelor	≥Bachelor	≥Bachelor
Background	NLP	Any	CS	CS	CS
First language	Chinese	non-English	Chinese	English	Chinese
Annotators	In-lab	External	Prolific	Prolific	Prolific
Human summary	4.70	4.60	4.65	4.70	4.68
PGN	2.92	1.53	1.60	1.90	1.88
HMNet	**3.52**	**2.68**	**2.23**	**2.90**	**3.08**
PGN(D_{KE})	3.20	$\underline{1.93}$	1.63	1.93	2.35
PGN(D_{RD})	3.15	1.90	$\underline{1.75}$	1.98	$\underline{2.53}$
PGN(D_{TS})	3.05	1.85	1.60	1.98	2.38
PGN(D_{ALL})	$\underline{3.33}$	1.85	1.65	$\underline{2.10}$	2.18
Fleiss' κ	0.48	0.19	0.20	0.13	0.05
Krippendorff's α		0.65	0.66	0.58	0.38

Table 1: The average informativeness values obtained in the original study and performed reproductions.

	Ranks of the final results					Averaged ranks			
	Original	ReproHum	R#1	R#2	R#3	ReproHum	R#1	R#2	R#3
Human summary	7	7	7	7	7	6.84	6.81	6.90	6.60
PGN	1	1	1.5	1	1	2.59	3.20	3.00	2.70
HMNet	**6**	**6**	**6**	**6**	**6**	**5.10**	**4.56**	**5.16**	**4.84**
PGN(D_{KE})	4	$\underline{5}$	3	2	3	$\underline{3.53}$	3.43	3.10	3.43
PGN(D_{RD})	3	4	$\underline{5}$	3.5	$\underline{5}$	3.48	$\underline{3.50}$	3.16	$\underline{3.81}$
PGN(D_{TS})	2	2.5	1.5	3.5	4	3.28	3.21	3.20	3.49
PGN(D_{ALL})	$\underline{5}$	2.5	4	$\underline{5}$	2	3.20	3.29	$\underline{3.48}$	3.14

Table 2: The ranked results of informativeness (the higher, the better) obtained in the original study (Original) and performed reproductions (ReproHum, R#1-R#3). We report both the ranks of the averaged informativeness from Table 1, as well as the ranks of informativeness averaged over all samples.

of whether the annotators have a background in computer science, are native English speakers or have a higher level of education, it seems that the main factor influencing this result is the participants' background in NLP (or in NLG tasks in particular) and potential prior experience with automatic summarizers.[5]

System ranking Among the automatically generated summaries, HMNet is consistently assessed as the best method for producing informative summaries and PGN as the worst. The PGN extensions are almost always all ranked in between the basic PGN and HMNet, but their ranking relative to each other varies greatly in different reproductions. This is because there are small absolute differences between them: the standard deviation of different PGN extensions' results is ≤0.14 for both the original study and all reproductions (even lower, at 0.04-0.07 for ReproHum, Repro #1 and Repro #2 reproductions).

Inter-annotator agreement The inter-annotator agreement is much lower in the reproduced studies as compared to the original experiment – Fleiss' κ is in the 0.1-0.2 range instead of the original 0.48. For Repro #3, κ is even lower. After looking at the correlation matrix between different annotators, we discovered that the responses of one annotator were poorly correlated with all the other annotators. We investigated the time spent on the annotation, but it was not different from the other annotators. The annotator also ranked human written summaries relatively higher than the other assessed summaries. Nevertheless, after excluding this annotator[6] the Fleiss' κ went up to 0.16, taking a value similar to that obtained in other reproductions.

[5] The original study was conducted in 2021, before the popularity of ChatGPT, which can also serve as a summarization engine. This may have raised annotators' expectations of the output quality of an AI-based system. However, we have no information on whether our annotators ever used ChatGPT for summarization.

[6] Repro #3 results recomputed for 3 annotators only are presented in Table 3).

Comparing different reproductions We do not observe very large differences that would indicate a significant impact of the factors influencing the selection of annotators. The reproduction that seems to stand out the most is Repro #3 (evaluation of all quality factors). As already mentioned, the informativeness of the PGN variants shows larger rating differences in this case (even when excluding the poorly correlated annotator).The absolute rating values are also consistently higher for all the methods, closer to the reproduced study. This may indicate that annotators responding to multiple quality criteria are more likely to try to split the overall quality rating into multiple factors than when presented with a single quality question. However, the observed differences against any other reproduction are not statistically significant.

Statistical analysis We performed a statistical analysis of the obtained informativeness results in all reproductions. Following the recommendations of Demšar (2006), we performed the non-parametric Friedman rank test with Nemenyi post-hoc analysis. For all reproductions, the null hypotheses of Friedman tests about the lack of differences in informativeness among all investigated summaries were rejected with low p-values ($p < 0.001$ for all reproductions). The results of the post-hoc analysis are presented in Figure 1 in the form of critical distance plots.

In all reproductions, the differences between the PGN baseline and all the PGN variants with additional annotations proposed by Feng et al. (2021) were not statistically significant at the $\alpha = 5\%$ significance level. In fact, the difference between human-written summaries and summaries provided by HMNet, the best automatic method, was also not significant due to the small sample size. In the main reproduction (ReproHum) and Repro #2 and #3, there is a statistically significant difference between HMNet and the PGN baseline. In contrast, the differences between HMNet and the enhanced variants of PGN are not statistically significant (except for PGN(D_{KE}) in Repro #2). In Repro #1, all automatic summarization methods are statistically indistinguishable.

Additional results from Repro #3 The results of Repro #3 include not only the informativeness values, but also the measurements of conciseness, coverage and the assessment of overall evaluation (the latter not being reported in the original work). The results are presented in Table 3. As mentioned earlier, the responses of one of the annotators were poorly correlated with those of the other three annotators, therefore we report the results averaged over all annotators (R#3) and the results averaged over three annotators only (R#3*). The discussion of the results will focus on the latter variant.

The general observation that our annotators evaluated all systems lower than in the original study remains true for coverage, but we obtained values of similar magnitude for conciseness. The ranking of the best performing methods resulting from the reproduction is similar to the original one for informativeness and coverage (Spearman correlations of 0.75 and 0.79, respectively) but differences are visible for conciseness (Spearman 0.39). Inter-annotator agreement is significantly lower than in the original study for all measures.

Looking at the overall binary quality evaluation, it seems that the PGN baseline is very weak, as none of the produced summaries were rated as good. The extensions of PGN improve the performance, but still fall significantly behind HMNet. Analysing the results of all measures, it seems that D_{RD} is the main cause of the improvement and combining it with other techniques (D_{ALL}) does not lead to further improvements, but, on the contrary, degrades the summaries.

5. Quantifying Reproducibility

Following the guidelines of the ReproHum shared task (Belz et al., 2023b, Sect. A5), we identify reproduction targets in the following categories:

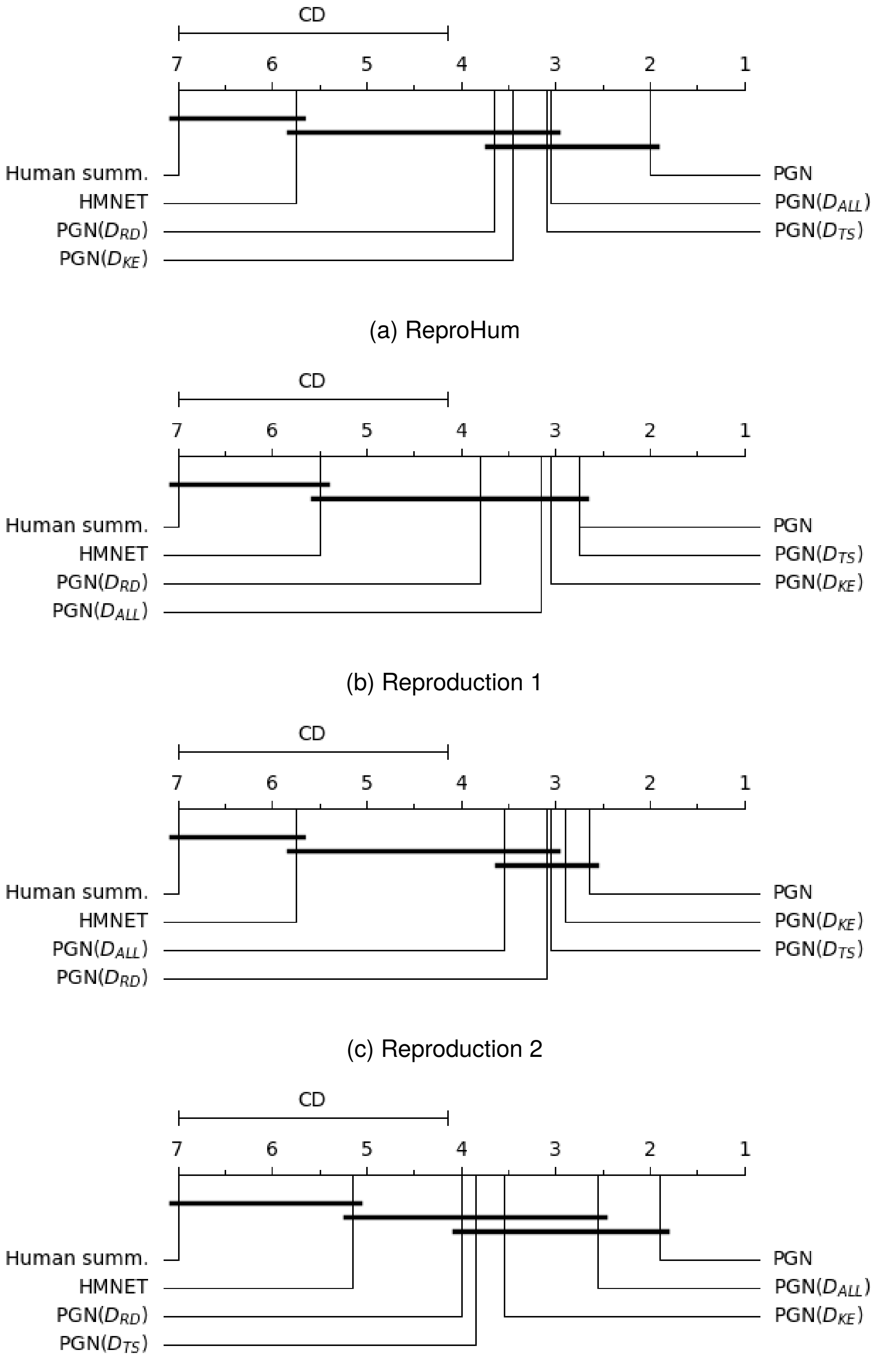
- Type I – numerical scores: the average informativeness of summaries generated by different methods
- Type II – sets of numerical values: the set of informativeness results for all the methods in the study

Type I Following the quantified reproducibility assessment by Belz et al. (2022), we computed the small sample coefficient of variation (CV*) as a measure of the degree of reproducibility for numerical scores. The results are given in Table 4.

The values of CV* computed for the original study and the main ReproHum reproduction are in the range of 48-63, except for the significantly lower values for HMNet and the summaries written by humans.

As to be expected, the coefficients of variation are smaller when computed for all the performed reproductions and the original study. Most CV* values are in the range of 28-33, again with the exceptions for HMNet and human summaries.

Type II results are evaluated with Pearson and Spearman correlation (Huidrom et al., 2022), as well as with the root-mean-square deviations from the original results. The results are presented in Table 5.



(a) ReproHum

(b) Reproduction 1

(c) Reproduction 2

(d) Reproduction 3

Figure 1: Critical distance diagrams showing the results of post-hoc Nemenyi tests performed for informativeness values obtained in the four performed reproductions. For all reproductions, the global Friedman test rejected the null hypothesis with $p < 0.001$. Critical distance plots present the average rank obtained in the Friedman test (the higher, the better) and show the difference between ranks that would imply statistical significance in the post-hoc analysis (critical distance). If the difference between the methods is not statistically significant, their results are connected with a thick horizontal line. More details on these plots can be found in (Demšar, 2006).

	Informativeness			Conciseness			Coverage			Overall	
	Orig.	R#3	R#3*	Orig.	R#3	R#3*	Orig.	R#3	R#3*	R#3	R#3*
Human summary	4.70	4.68	4.97	3.85	4.28	4.43	4.35	4.58	4.80	0.98	1.00
PGN	2.92	1.88	1.53	3.08	2.55	2.27	2.70	2.05	1.67	0.13	0.00
HMNet	**3.52**	**3.08**	**2.80**	2.40	3.00	2.97	**3.40**	**3.18**	**3.00**	**0.50**	**0.40**
PGN(D_{KE})	3.20	2.35	2.27	3.08	**3.23**	**3.10**	3.00	2.33	2.10	0.13	0.07
PGN(D_{RD})	3.15	2.53	2.53	**3.25**	3.18	**3.10**	3.00	2.53	2.53	0.13	0.10
PGN(D_{TS})	3.05	2.38	2.17	3.10	3.03	2.87	3.17	2.33	2.13	0.11	0.04
PGN(D_{ALL})	3.33	2.18	1.90	**3.25**	2.85	2.70	3.10	2.08	1.80	0.10	0.10
Fleiss' κ	0.48	0.05	0.16	0.40	0.01	0.03	0.41	0.03	0.11	0.47	0.61
Krippendorff's α		0.38	0.51		0.13	0.15		0.35	0.45	0.47	0.61

Table 3: The average informativeness, conciseness, coverage and overall binary evaluation of summaries as obtained in the original human evaluation (Orig.) and our Repro #3 (R#3). Additionally, we also report reproduction results computed on data from 3 annotators only (R#3*) - see more details in the text.

CV*	ReproHum	All Repro.
Human summary	2.14	1.01
PGN	62.28	31.71
HMNet	27.02	18.51
PGN(D_{KE})	49.36	30.91
PGN(D_{RD})	49.36	28.51
PGN(D_{TS})	48.83	29.11
PGN(D_{ALL})	56.97	32.86

Table 4: The small-sample coefficient of variation (CV*) of informativeness computed for original and ReproHum study (2 samples) and for all the reproductions (5 samples).

The Pearson correlations are very high for all the reproduction studies, which can be attributed to the fact that the human summary scores are relatively high outliers in all the studies (after removing them, the correlations drop from 0.97-0.99 to 0.78-0.88). This is also reflected in the lower Spearman correlations, which are more robust to outliers. The lowest Spearman correlation was obtained for Reproduction 3 (0.68) which is the only correlation in this study that is not statistically significant ($\alpha = 5\%$).Note that the sample size is very small (7).

Finally, RMSE values of around 1 reflect the general tendency of our annotators to rate automatic summaries lower than in the original study. The closest results to the original study, as measures by RMSE, were obtained in the Reproduction 3 where all quality factors were evaluated.

6. Summary

From the results of the original study, the authors draw three major conclusions (see Sec. 4.5 in Feng et al., 2021):

1. "HMNet gets the best score in informativeness and coverage", which was confirmed by our reproductions.

2. "Our method can achieve higher scores in all three metrics", which again is in line with the results of our reproductions.

3. "We also find there is still a gap between the scores of generated summaries and the scores of golden summaries" – which was not only confirmed in our reproductions, but also the gap seems substantially larger than in the original study.

	Pearson	Spearman	RMSE
ReproHum	0.99	0.85	1.16
Repro #1	0.98	0.88	1.35
Repro #2	0.98	0.88	1.00
Repro #3	0.97	0.68	0.77

Table 5: The values of root-mean-square deviation, Pearson and Spearman correlations computed between the original and reproduced results.

Nevertheless, the results of the original study also provided evidence that the combination of all proposed annotations (D_{ALL}) gives the best informativeness among the PGN variants and that the gap against the better performing HMNet is relatively small (0.19). This was not confirmed by our reproductions. D_{ALL} was the worst PGN extension evaluated in two reproductions, and the best and second best in the other two reproductions. Similarly, the reported gap between the best PGN extension and HMNet ranged from 0.48 to 0.8 on a 5-point scale, at least two and a half times larger than in the original study.

7. Discussion

We can attempt to draw some conclusions from the analysis of the differences between our reproductions: Firstly, mother tongue, level of education or field of study do not seem to have a significant impact on the results of human evaluation in the summarisation task. The only exception is a very specific background in NLP technologies. Second, when working on reproduction experiments, it might be better to evaluate all quality factors, even if were are interested in reproducing the result for a single quality factor in particular. Finally, we believe that it is always helpful to carry out a statistical analysis of the results obtained. Even if the analysis is not conclusive, e.g. due to the small sample size, it gives a much better picture of the variability of the results and the conclusions that can be drawn from them.

8. Acknowledgements

This research was supported by the European Research Council (Grant agreement No. 101039303 NG-NLG) and by Charles University projects GAUK 40222 and SVV 260 698. It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

9. Bibliographical References

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, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. 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, Aberdeen, Scotland, UK. 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. INCOMA Ltd., Shoumen, Bulgaria.

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

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, 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. 2023a. 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.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Huiyuan Lai, Chris van der Lee, Emiel van Miltenburg, Yiru Li, Saad Mahamood, Margot Mieskes, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Pablo Mosteiro Romero, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023b. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp.

Marc Brysbaert. 2019. How many words do we read per minute? a review and meta-analysis of reading rate.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. 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 1479–1491, Online. Association for Computational Linguistics.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard.

Rudali Huidrom, Ondřej Dušek, Zdeněk Kasner, Thiago Castro Ferreira, and Anya Belz. 2022. Two reproductions of a human-assessed comparative evaluation of a semantic error detection system. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 52–61, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. 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.

Zixiu Wu, Simone Balloccu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Are experts needed? on human evaluation of counselling reflection generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6906–6930, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

10. Language Resource References

Carletta, Jean and Ashby, Simone and Bourban, Sebastien and Flynn, Mike and Guillemot, Mael and Hain, Thomas and Kadlec, Jaroslav and Karaiskos, Vasilis and Kraaij, Wessel and Kronenthal, Melissa and Lathoud, Guillaume and Lincoln, Mike and Lisowska, Agnes and McCowan, Iain and Post, Wilfried and Reidsma, Dennis and Wellner, Pierre. 2006. *The AMI Meeting Corpus: A Pre-announcement*. Springer Berlin Heidelberg.

Gliwa, Bogdan and Mochol, Iwona and Biesek, Maciej and Wawer, Aleksander. 2019. *SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization*. Association for Computational Linguistics.

A. Human Evaluation Datasheet (HEDS)

Human Evaluation Datasheet (HEDS, Shimorina and Belz, 2022) for the main ReproHum reproduction (see Sec. 3.1) is provided in the ReproHum GitHub repository.[7]

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