TUDA-Reproducibility @ ReproGen: Replicability of Human Evaluation of Text-to-Text and Concept-to-Text Generation

Christian Richter, Yanran Chen, Steffen Eger
Computer Science Department
Technical University of Darmstadt (TUDA)
chrisrichter145@gmail.com, chenyr1996@hotmail.com, eger@aiphes.tu-darmstadt.de

Abstract

This paper describes our contribution to the Shared Task *ReproGen* by Belz et al. (2021), which investigates the reproducibility of human evaluations in the context of Natural Language Generation. We selected the paper "Generation of Company descriptions using concept-to-text and text-to-text deep models: data set collection and systems evaluation" (Qader et al., 2018) and aimed to replicate, as closely to the original as possible, the human evaluation and the subsequent comparison between the human judgements and the automatic evaluation metrics. Here, we first outline the text generation task of the paper of Qader et al. (2018). Then, we document how we approached our replication of the paper's human evaluation. We also discuss the difficulties we encountered and which information was missing. Our replication has medium to strong correlation (0.66 Spearman overall) with the original results of Qader et al. (2018), but due to the missing information about how Qader et al. (2018) compared the human judgements with the metric scores, we have refrained from reproducing this comparison.

1 Introduction

Reproducibility is an utmost priority in research to ensure reliability of scientific findings. Informally, it describes the ability to repeat a study, beginning with the same starting point, using the same resources (if possible) and achieving the same results and conclusions (Pineau et al., 2020). Reproducibility requires that approaches in publications be recorded in such a way that previously uninvolved parties can comprehend and recreate them (Fokkens et al., 2013). However, reproducibility is a complex requirement which often fails because of missing details (like not described data sets or missing key parameters)—such aspects, even though they may appear minor at first sight, either prevent reproducibility altogether or at least distort the results (Raff, 2019; Wieling et al., 2018). One reason for such failures of reproducibility may be lack of widely accepted definitions and practical conceptualization of reproducibility, as there is currently no consensus on how and to what level of detail research should be documented (Cohen et al., 2018).

The Shared Task *ReproGen* (Belz et al., 2021) deals with the reproducibility problem. In particular, it aims to investigate reproducibility of human evaluation. The findings of ReproGen should yield general insights into how reproducibility can be improved. The task in ReproGen is to replicate either one of the pre-selected studies or a self-selected study from the field of Natural Language Generation (NLG) and to document the findings.

In this paper, we report on our reproducibility of the work "Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation" (`CompDesc` for short) by Qader et al. (2018). This work analyzes multiple sequence-to-sequence models that were used to generate short company descriptions from Wikipedia articles. This includes both automatic and human evaluation which are then compared with each other. Our replication focuses on the human evaluation, in accordance with the general outline of ReproGen.

2 `CompDesc` and our replication

We first describe the paper `CompDesc`, then outline how we replicated its human evaluation. Finally, we compare both experiments.

2.1 The paper `CompDesc`

The paper `CompDesc` first creates a data set of Wikipedia articles about companies[1]. Then, using four concept-to-text and two text-to-text approaches, they generate short summaries out of

[1] https://gricad-gitlab.univ-grenoble-alpes.fr/getalp/wikipediacompanycorpus

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 301–307,
Aberdeen, Scotland, UK, 20-24 September 2021.

this data. Figure 1 shows an example from our experiment, which is what the evaluators can see during the evaluation. The title and the description at the top as well as the info box at the right margin, which are typically present in every Wikipedia article, serve as input. The language generation models then generate the summary either from the description or the info box, depending on the type of the text generation system. Afterwards, Qader et al. (2018) evaluated the system performance on the test set of their Wikipedia company corpus using five automatic evaluation metrics. Table 7 in Appendix A.3 shows the results of the automated evaluation. In addition to that, they conduct a human evaluation using a selection of randomly sampled summaries with 19 test persons where each one evaluated 10 summaries. But the human evaluators did not know that some of the summaries were actually human generated, namely the references. For that, the humans assessed the criteria *information coverage*, *information redundancy*, *semantic adequacy* and *grammatical correctness* on a 5-point Likert scale. Finally, Qader et al. (2018) compared the results of the two evaluation methods.

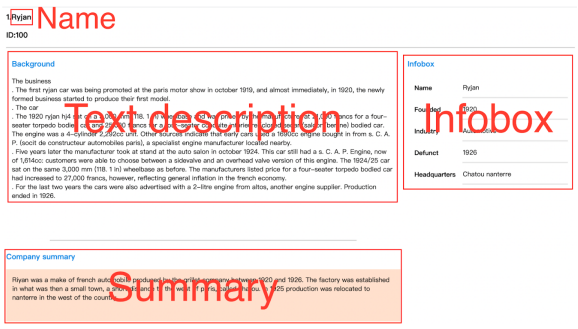

Figure 1: Example of Human Evaluation: the summary is created from the other fields. All 4 boxes are presented during the human evaluation process.

2.2 Replication of `CompDesc`

There were two phases in our replication study: 1) Preparation, where the goals of reproduction and needed resources were determined; 2) the human evaluation experiment, where we collected the human ratings, that were then compared to the original results.

Preparation In the preparation phase, there were three resources initially provided by Belz et al. (2021) as part of the shared task (see Appendix A.1), namely 1) the original paper (Qader et al., 2018), which describes the implementation as well as the methods and data used; 2) an incomplete human evaluation data sheet filled out by the authors of (Qader et al., 2018), which should also be filled out by the participants of the shared task later; 3) a link to a GitLab repository that contains code for a web-based survey tool called "FlexEval" (Fayet et al., 2020). The original code was not available, also not upon request.

Based on the information and resources available, we first identified which results should be replicated: **the average scores of the human evaluation based on a 5-point Likert scale per system** (see Table 1), and, as a secondary goal, **the comparison of human and automatic evaluation metrics using Spearman's correlation** (see Table 4). Then, we determined the resources needed to reproduce the human evaluation, which include the system outputs and references, the data and ideally the code for computing the correlations. However, none of the above was included in the Shared Task resources. Upon request, the authors provided us with parts of the data, including the summaries they used to conduct the human evaluation (both as CSV and HTML files) and a CSV file containing their human evaluation scores, whose reproduction is the primary goal of this report.

Human Evaluation Experiment In order to keep our reproduction as close as possible to the original in terms of content and appearance, the identical data sets were selected for reproduction using the provided HTML files.

In the beginning, a unique identification number was assigned to each summary to match the results to the corresponding summaries. After that, 19 files, each containing 10 summaries, were randomly created out of the original files. In addition, a survey was created using Google Form[2] to collect the evaluator ratings of the four criteria, each on the basis of a 5-point Likert Scale. 19 people from the authors' social environment volunteered as participants for this study. They were not English native speakers, similar to the participants of `CompDesc`. However, `CompDesc` does not explain why these conditions were chosen. This may not have been intentional, but a result of the composition of the participants. We have decided to take this into account anyway. When conducting the human evaluation, each participant was given one of the 19 HTML files and a link to the Google

[2]https://www.google.de/intl/en/forms/about/

Form via E-Mail or other chat apps (see Appendix A.2).

After obtaining the human ratings, we exported the data using the same format as the original one. We calculated the average scores directly based on this file. But for reproducing the correlation matrix, some further resources were needed. Since the paper is ambiguous about how the results were computed and the corresponding code and data were missing, we tested different calculation approaches to determine the original calculation. Unfortunately, it didn't succeed in the end. We could only reproduce a part of the correlation values on the basis of the original human evaluation results that Qader et al. (2018) provided us. Table 5 presents this special case, which we will describe in detail in Section 3. In the end, we examined the similarities between the original and reproduced results.

2.3 Assessment

Comparing with the original experiment, there are several notable differences. First, the original study used "FlexEval" (Fayet et al., 2020) to conduct the survey, which probably showed the evaluation data and corresponding questions side by side in a web application and the evaluators can answer the questions by scrolling down. In their paper, Qader et al. (2018) only stated that they "set up a web-based experiment" (Qader et al., 2018), but they did not mention what tool they used. However, since the tool is very complex to configure and adequate guidance was not available, we used Google Forms[3] instead. However, we made sure that the participants received the same data presentation.

The use of "FlexEval" only became apparent with additional information from the shared task, as the authors mentioned it in the human evaluation data sheet. But in our survey, the presentation of the data and the input mask were accessible through two separate sources. In contrast to the participants of the original study, who were all members of a lab, the participants of the replication were only selected based on their connection to us.

Besides those distinct differences from the original experiment, we made several assumptions because of the inaccuracies and the missing information found during the preparation, which could influence possible deviations of the results between original and replication. We describe these in the following:

[3]https://www.google.de/intl/en/forms/about/

1) There is an inconsistency in the description of the experiments sets. Qader et al. (2018) stated in their paper that each of the 19 participants evaluated 10 summaries, resulting in a total number of 190. However, it was also stated in the paper that 30 summaries were evaluated for all 7 systems (including reference), which makes a total of 210 summaries. When asked, the authors explained that a random selection was made from the 210 summaries. This agrees with the raw human evaluation results we received on request. Therefore, we relied on the explicit specification of 19 times 10 random summaries.

2) Qader et al. (2018) perform a manual quality checking of the results of the human evaluation, but do not go into detail about the procedure. To be able to guarantee a minimum of quality, we considered an evaluation invalid when the majority of the answers were illogical. This occurred only once, where a participant randomly selected the values 3 and 4 independently of the summary quality. In this case, we passed the task to an additional participant for re-evaluation.

Nevertheless, the replication follows the original in the essential points such as the requirements for evaluators, the number of evaluators, the amount of evaluated items, the identical set of questions, the format of data, and the survey guidelines which prohibit to ask questions during the experiment. Therefore, we conclude that, assuming the same basic conditions, a comparison of the results below is legitimate.

3 Results

Table 1 displays the human evaluation results of Qader et al. (2018), whereas Table 2 shows our replicated results. As one can see, different levels of variation show up between the two experiments. Larger deviations of more than one point can only be seen twice, all other deviations are smaller. These deviations may have been caused by various factors. In general, smaller differences are always possible in stochastic environments. It also cannot be ruled out that the differences may result from minor but recurring discrepancies of the score as well as the participants in the two studies could have rated the results fundamentally differently, but with a simple average deviation of 0.47, which is only 14% off the average value.

In addition, we calculated Spearman's ρ and Pearson's r correlations between the values in the

	cover.	non-redun	semant.	gramm.
Reference	3.1	4.6	3.9	4.2
C2T	2.9	2.9	3.3	3.6
C2T_char	2.3	3.9	2.8	3.0
C2T+pg	2.3	4.5	4.0	4.3
C2T+pg+cv	2.7	3.9	3.6	4.2
T2T+pg	1.8	3.3	2.9	3.7
T2T+pg+cv	2.3	3.8	2.4	3.5

Table 1: ORIGINAL: The original human evaluation results taken from Qader et al. (2018).

	cover.	non-redun	semant.	gramm.
Reference	3.9	4.1	3.9	4.0
C2T	2.5	3.8	2.6	3.2
C2T_char	3.0	**2.8**	3.1	3.5
C2T+pg	2.6	4.2	**2.9**	3.8
C2T+pg+cv	3.0	4.1	3.9	4.1
T2T+pg	2.6	3.5	2.7	4.0
T2T+pg+cv	2.9	4.1	2.8	4.4

Table 2: REPLICATION: The replication results of the human evaluation. Differences of more than 1 are bold.

two tables on each axis. From Table 3, we observe that the reproduced evaluation scores for the systems C2T+pg, C2T+pg, T2T+pg and T2T+pg+cv are highly correlated with the original values, but this may be unreliable due to the small number of input values. Unfortunately, we were not able to compare the scores at the summary-level, because of the missing information about the arrangement in the original experiment. However, if we calculate a single correlation using both methods between all values of both tables, we get a more reliable score. The values of 0.66 respectively 0.7 represent a moderate to strong statistical significant correlation (Taylor, 1990; Schober et al., 2018).

	ρ	r
All	0.66*	0.70*
Reference	0.95	0.82
C2T	0.20	-0.05
C2T_char	0.20	-0.32
C2T+pg	1.0*	0.83
C2T+pg+cv	0.80	0.97*
T2T+pg	1.0*	0.86
T2T+pg+cv	0.60	0.95*
cover.	0.41	0.58
non-redun.	0.64	0.38
semant.	0.36	0.54
gramm.	0.14	0.33

Table 3: Spearman's ρ and Pearson's r correlations between Table 1 and Table 2. Values marked with * show a significant correlation ($p \leq .05$).

To figure out how Qader et al. (2018) computed the correlations specifically, we conducted some further experiments based on the original data from Qader et al. (2018), which contains the human judgements for each summary. In the first step, we proved the validity of the data by a successful reproduction of the values in Table 1. Afterwards, we made several attempts regarding the source of the metric scores, the level at which the correlations were computed and whether the correlated values included the scores for the references, to achieve a valid reproduction of the original correlation matrix, using only the original data.

After that, despite not discovering the original setup, there is one noteworthy case (see Table 5) where we successfully reproduced the correlations between the results of the 5 automatic evaluation metrics and that between the human judgements of the 4 criteria (values outside the black square). Surprisingly, a large gap still exists for the comparison between the metric scores and the human scores (values in the black square). We can draw a completely different conclusion from these reproduced correlations. E.g., Table 5 shows that METEOR, ROUGE-L, and CIDEr are highly correlated with redundancy (green marker), but in Table 4, which displays the results of the original paper (Qader et al., 2018), there is no significant correlation between redundancy and any metric at all. Considering that Qader et al. (2018) explicitly stated in the paper that the references were excluded when comparing the metric scores with the human judgements, we also computed the correlations once without the references. However, this attempt only led to a worse result (see Table 6), since none of the correlation values could be reproduced.

Since Qader et al. (2018) were not able to provide us with the original code or the corresponding information, it was impossible to determine the reason for the difference. For this reason and the consequent non-comparability of the results, we have refrained from reproducing the correlation matrix using the human evaluation results obtained in this replication study.

4 Conclusion

In this replication, we could not reproduce all results of the original study "Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evalu-

	BLEU	NIST	METEOR	ROUGE-L	CIDEr	Coverage	Redundancy	Semantic
NIST	0.64							
METEOR	0.29	0.00						
ROUGE-L	0.11	-0.07	0.96					
CIDEr	0.00	-0.14	0.86	0.93				
Coverage	0.64	0.21	0.64	0.50	0.39			
Redundancy	-0.14	0.00	0.36	0.54	0.64	0.25		
Semantic	-0.14	-0.43	0.71	0.82	0.93	0.25	0.64	
Grammar	-0.38	-0.38	0.59	0.74	0.90	0.00	0.61	0.92

Table 4: ORIGINAL: Correlation matrix from Qader et al. (2018), human vs. automatic metric correlations are in the black square. Color markers indicate significant correlations, the different colors are for better readability

	BLEU	NIST	METEOR	ROUGE-L	CIDEr	Coverage	Redundancy	Semantic
NIST	0.64							
METEOR	0.29	0.00						
ROUGE-L	0.11	-0.07	0.96					
CIDEr	0.00	-0.14	0.86	0.93				
Coverage	0.57	0.28	0.29	0.21	0.18			
Redundancy	-0.04	-0.11	0.89	0.96	0.96	0.25		
Semantic	-0.21	-0.36	0.36	0.54	0.71	0.25	0.64	
Grammar	-0.45	-0.31	0.23	0.45	0.68	0.00	0.61	0.92

Table 5: REPLICATION: Correlation matrix reproduced based on the human evaluation results from Qader et al. (2018), computed at the system-level (**including** reference), using automatic metric scores from Table 7 in Appendix A.3. Color markers indicate significant correlations, the different colors are for better readability.

	BLEU	NIST	METEOR	ROUGE-L	CIDEr	Coverage	Redundancy	Semantic
NIST	0.43							
METEOR	-0.14	-0.60						
ROUGE-L	-0.43	-0.71	0.94					
CIDEr	-0.6	-0.83	0.77	0.89				
Coverage	0.31	-0.14	-0.14	-0.26	-0.31			
Redundancy	-0.66	-0.77	0.83	0.94	0.94	-0.20		
Semantic	-0.60	-0.82	0.09	0.31	0.60	0.09	0.49	
Grammar	-0.82	-0.60	-0.03	0.26	0.60	-0.25	0.49	0.89

Table 6: REPLICATION: Correlation matrix reproduced based on the human evaluation results from Qader et al. (2018), computed at the system-level (**excluding** reference), using automatic metric scores from Table 7 in Appendix A.3. Color markers indicate significant correlations, the different colors are for better readability.

ation" of Qader et al. (2018)

The primary goal of ReproGen (Belz et al., 2021) was to conduct an equivalent human evaluation with the aim of obtaining comparable values. We were able to reproduce the human evaluation and obtain results that are not only apparently comparable but also to obtain a moderate to strong statistical significant correlation (Taylor, 1990; Schober et al., 2018) using both Spearman's ρ and Pearson's r. However, this has taken a lot of time to gather all the information needed from both the paper and the authors.

In contrast to the first one, our secondary objective, namely to investigate whether we could obtain comparable inferences with the reproduced correlation matrix based on our human evaluation results, was not successful. We had to make several assumptions of missing information and even with that, we were not even able to recalculate the original results by using the human evaluation results from Qader et al. (2018). Therefore, we have refrained from a comparison with our data.

References

Anya Belz, Shubham Agarwal, Anastasia Shimorina, 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*. Website of the shared Task: https://reprogen.github.io/, last accessed: August 30, 2021.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. 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.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Cédric Fayet, Alexis Blond, Grégoire Coulombel, Claude Simon, Damien Lolive, Gwénolé Lecorvé, Jonathan Chevelu, and Sébastien Le Maguer. 2020. FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. In *6e*

conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition), pages 22–25, Nancy, France. ATALA.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria. Association for Computational Linguistics.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2020. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program).

Raheel Qader, Khoder Jneid, François Portet, and Cyril Labbé. 2018. Generation of company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 254–263, Tilburg University, The Netherlands. Association for Computational Linguistics.

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems*, 32:5485–5495.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Richard Taylor. 1990. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Squib: Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.

A Appendix

A.1 Resources

This section lists the external resources that were used and describes whether they were made available in advance or have been collected during the implementation.

- The original paper (Qader et al., 2018), included by the task of ReproGen (Belz et al., 2021).
- The Human Evaluation Datasheet v1.0 [4] (Belz et al., 2020) filled out by Qader et al. (2018) (incomplete), included by the task of ReproGen (Belz et al., 2021).
- The web based evaluation tool "FlexEval" (Fayet et al., 2020), included by the task of ReproGen (Belz et al., 2021).
- Google Forms[5], used to do the survey.
- The part of the Wikipedia data sets that was used for the human evaluation, provided upon request by Qader et al. (2018).
- The anonymized original human evaluation results, provided upon request by Qader et al. (2018).

A.2 Access to Resources

The data we are able to publish, including code and results, is available in this Github Repository: https://github.com/der-Richter/TUDA-Reproducibility-ReproGen. To obtain access to the data from the original study, please contact Qader et al. (2018) directly.

A.3 Tables

System	BLEU	NIST	METEOR	ROUGE_L	CIDEr
C2T	0.0608	1.9322	0.0906	0.2092	0.1872
C2T_char	0.0750	1.0975	0.1159	0.2665	0.2731
C2T+pg	0.0413	0.0893	0.1076	0.2668	0.2836
C2T+pg+cv	0.0490	0.2349	0.1045	0.2589	0.2734
T2T+pg	0.0567	1.9690	0.1002	0.2212	0.1992
T2T+pg+cv	0.0558	2.1188	0.1024	0.2216	0.1974

Table 7: System results on the test set of the Wikipedia Company Corpus from Qader et al. (2018)

[4] https://drive.google.com/file/d/1_74CJ_n8vSPm8FvA6P_Sg49aZp3kecRo/view

[5] https://forms.gle/AQJS2s2GAHKAKPgd9