ReproHum #0729-04: Human Evaluation Reproduction Report for "MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes"

Simeon Junker
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
simeon.junker@uni-bielefeld.de

Abstract

Human evaluation is indispensable in natural language processing (NLP), as automatic metrics are known to not always align well with human judgments. However, the reproducibility of human evaluations can be problematic since results are susceptible to many factors, the details of which are often missing from the respective works. As part of the ReproHum project, this work aims to reproduce the human evaluation of a single criterion in the paper "MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes" (Gu et al., 2022). The results of our reproduction differ noticeably from those of the original study. To explain this discrepancy, we discuss unavoidable differences in the experimental setup, as well as more general characteristics of the selected domain and the generated summaries.

1 Introduction

Human evaluation is generally considered the gold standard in NLP research (Belz et al., 2020). While automatic metrics are usually easier and cheaper to use, they have been shown as problematic in different ways: For example, standard metrics are often used in inappropriate settings and without reporting important details such as version information, and they do not always correlate well with human judgments (Belz and Reiter, 2006; Novikova et al., 2017; van der Lee et al., 2019; Sai et al., 2022; Chen et al., 2022; Schmidtova et al., 2024).

Human evaluations can solve some of those issues, but come with their own challenges. Apart from higher costs and time expenditures, it has been shown that human evaluations in existing research do not always rely on the same terminology (Belz et al., 2020) and that the evaluation outcomes can be affected by a multitude of parameters, the details of which are often missing from reports (Howcroft et al., 2020; Belz et al., 2023). As a consequence, *reproducibility* is a core issue for human evaluation, potentially casting doubt on the validity of reported results and conclusions (Belz et al., 2021).

Against this background, the ReproHum project and associated ReproNLP shared task (Belz et al., 2025) aim to systematically test the reproducibility of human evaluations and strengthen transparency and reliability in NLP research. As part of this project, we attempt to reproduce the human evaluation in the paper "MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes" (Gu et al., 2022).

In the following, we first outline the content of Gu et al. (2022)'s work. After that, we describe the details of the human evaluation carried out in this study and the differences from the original work. Finally, we compare the results of our reproduction study with the original findings.

2 Original Study

In their work, Gu et al. (2022) look at extractive summarization, i.e., selecting a subset of sentences from a source document which adequately summarize the content of the full text. For this, the authors propose *MemSum*, an extractive summarizer based on reinforcement learning that is designed for long documents. MemSum utilizes a multi-step episodic Markov decision process to iteratively select sentences from the source document. For each decision, the system considers a broad set of information, i.e., the content of the current sentence, the global context of the rest of the document, and the sentences selected in previous steps.

The system is tested with ROUGE (Lin, 2004) as an automatic metric, showing state-of-the-art performance on long document datasets such as PubMed, arXiv (Cohan et al., 2018), and GovReport (Huang et al., 2021).

In addition to this, Gu et al. (2022) conduct a

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 569–583
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

	Original	Reproduction
Quality Criterion	overall quality, coverage, non-redundancy	overall quality
Number of Items	63	63
Number of Systems	2	2
Number of Participants	4	4
Participants/Item	1	4
Compensation	unknown	13.98 – 16.25 € / h
Gender Split	unknown	1 female, 3 male
Professional Status	Master's / PhD Students (Computer Science)	Bachelor's / Master's Students (Computational Linguistics)
English proficiency	unknown	fluent, second language

Table 1: Comparison between the human evaluation in the original work (Gu et al., 2022) and our reproduction.

human evaluation where MemSum is compared to a strong baseline, i.e., the existing NeuSum (Zhou et al., 2018) summarizer. The evaluation is divided into two parts, where MemSum generates summaries with adaptive length (Experiment I) or is fixed to a number of 7 sentences (Experiment II). In both experiments, evaluators rate summaries for texts from the PubMed dataset which are generated by MemSum and NeuSum, respectively. The generated summaries are compared to ground-truth abstracts written by humans with respect to *coverage*, *non-redundancy* and *overall quality*. The results show that NeuSum achieves slightly better coverage, but MemSum summaries are rated significantly higher for non-redundancy. MemSum also exceeds the baseline for overall quality, although this difference is not statistically significant.

In this work, we aim to reproduce Experiment II in Gu et al. (2022), focusing on the *overall quality* criterion and disregarding *coverage* and *non-redundancy*. With regard to this scope, the main finding of the original study can be summarized as follows: **MemSum generates summaries of higher overall quality than the NeuSum baseline.**

3 Method

In our evaluation setup, we tried to follow the procedure of Gu et al. (2022) as closely as possible. Our code is available on GitHub[1]. More details can be found in our Human Evaluation Data Sheet (HEDS, Shimorina and Belz 2022; Belz and Thomson 2024)[2].

3.1 Material

For Experiment II in their human evaluation, Gu et al. (2022) sampled 63 documents from the test set of the PubMed dataset. For each document, they retrieved a ground-truth abstract as a reference summary and generated two summaries with MemSum and NeuSum, respectively. We use the same items as in the original evaluation.

3.2 Evaluators

We recruited four evaluators (one female, three male). At the time of the experiment, all evaluators were students in Computational Linguistics and related fields and employed as student assistants in our group. The participants were paid by the hour according to the local statutory rate (13.98 € or 16.25 € / hour, depending on educational attainment). All participants are native German speakers who are proficient in English as a second language.

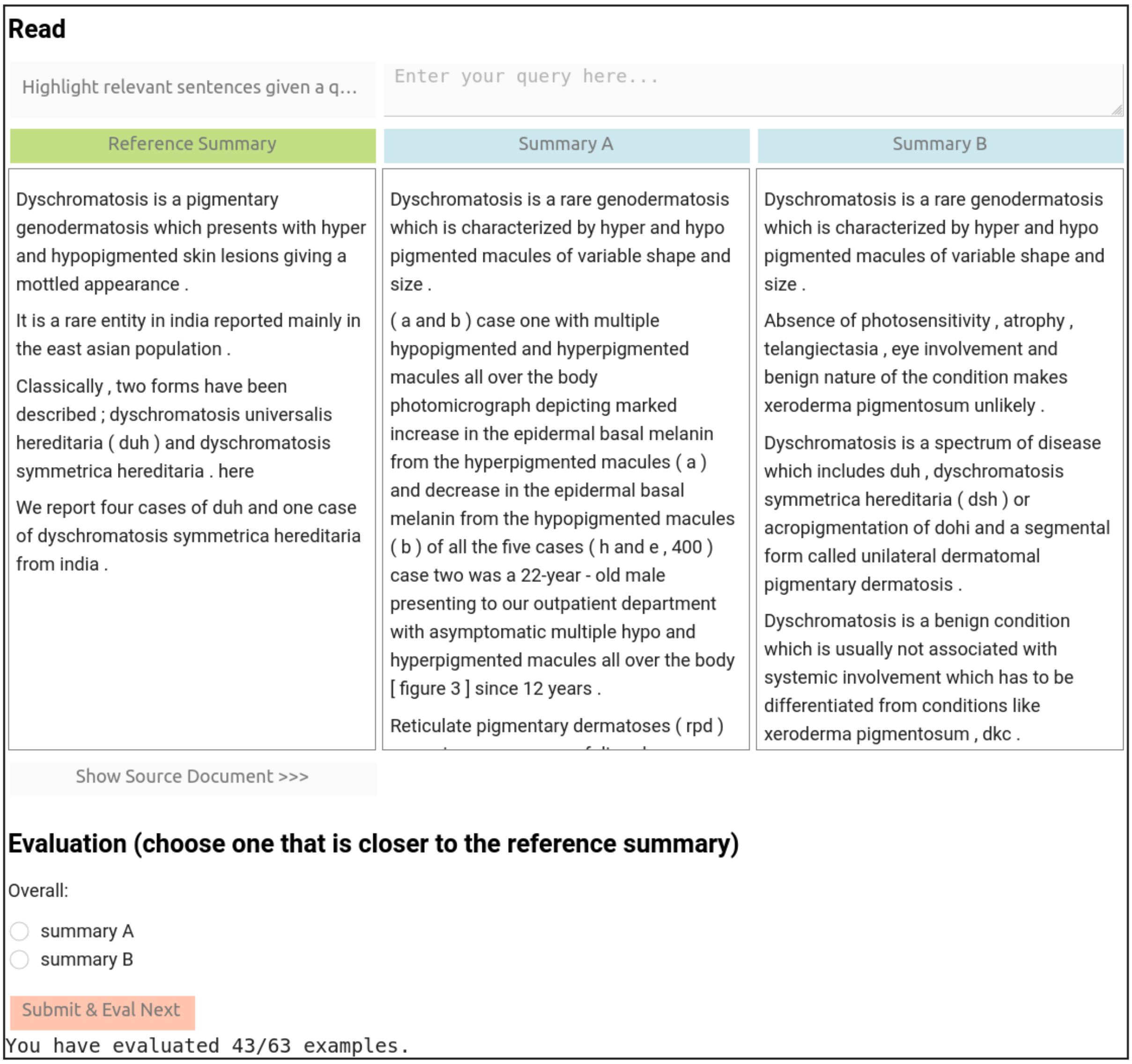
3.3 Evaluation Procedure

We asked our evaluators to rank the summaries of the two systems according to their quality. We did not provide further guidelines, but relied on the short instructions included in the evaluation notebook published by the original authors.

The evaluation was carried out through an interactive web interface in Google Colab, which was based on the published evaluation notebook of the original project (see the screenshot in Figure 1). We note, however, that in contrast to the interface reported in the original paper, our notebook did not include a function for skipping items (see Section 3.4). For each document, the interface presented a reference summary next to two generated summaries in random order, one generated by Mem-

[1] github.com/clause-bielefeld/ReproHum0729-04

[2] github.com/nlp-heds/repronlp2025



Reference Summary

Dyschromatosis is a pigmentary genodermatosis which presents with hyper and hypopigmented skin lesions giving a mottled appearance .

It is a rare entity in india reported mainly in the east asian population .

Classically , two forms have been described ; dyschromatosis universalis hereditaria (duh) and dyschromatosis symmetrica hereditaria . here

We report four cases of duh and one case of dyschromatosis symmetrica hereditaria from india .

Summary A

Dyschromatosis is a rare genodermatosis which is characterized by hyper and hypo pigmented macules of variable shape and size .

(a and b) case one with multiple hypopigmented and hyperpigmented macules all over the body photomicrograph depicting marked increase in the epidermal basal melanin from the hyperpigmented macules (a) and decrease in the epidermal basal melanin from the hypopigmented macules (b) of all the five cases (h and e , 400) case two was a 22-year - old male presenting to our outpatient department with asymptomatic multiple hypo and hyperpigmented macules all over the body [figure 3] since 12 years .

Reticulate pigmentary dermatoses (rpd)

Summary B

Dyschromatosis is a rare genodermatosis which is characterized by hyper and hypo pigmented macules of variable shape and size .

Absence of photosensitivity , atrophy , telangiectasia , eye involvement and benign nature of the condition makes xeroderma pigmentosum unlikely .

Dyschromatosis is a spectrum of disease which includes duh , dyschromatosis symmetrica hereditaria (dsh) or acropigmentation of dohi and a segmental form called unilateral dermatomal pigmentary dermatosis .

Dyschromatosis is a benign condition which is usually not associated with systemic involvement which has to be differentiated from conditions like xeroderma pigmentosum , dkc .

Evaluation (choose one that is closer to the reference summary)

Overall:

○ summary A
○ summary B

```
You have evaluated 43/63 examples.
```

Figure 1: Screenshot of the evaluation interface as used in our work.

Sum and the other by the baseline NeuSum system. Using HTML radio buttons, the evaluators should indicate which of the two generated summaries has a higher overall quality or is more consistent with the reference summary. For additional assistance, the interface included a highlight function that marks text spans in color that correspond to the content of an input query. Sent2vec sentence embeddings (Pagliardini et al., 2018) were used to determine the relevance of text passages.

For each item, the system rated as better is ranked #1, while the other is ranked #2. As in the original paper, we tested the item pairs for instances where both systems gave the exact same response and replaced the evaluator ratings with rank #1 for both systems in those cases. In our results section we report the mean ranks per system, averaged over all items and evaluators.

3.4 Known Differences to Original Study

Our study differs from the original study in some aspects. A summary of the comparison between the original study and our reproduction can be seen in Table 1.

First, as described in Section 2, the original evaluation is not restricted to the *overall quality* criterion, but also includes the *coverage* and *non-redundancy* of the extracted summaries. We focus on overall quality, excluding the other criteria from the interface.

Second, the authors report a function in the interface to skip items where no clear decisions can be made. This function was not available in the published evaluation notebook and is therefore not included in our reproduction, i.e., evaluators must decide on a ranking for all items.

Finally, Gu et al. (2022) do not provide details regarding the gender and language skills of the evaluators or the compensation for the experiments. Additionally, the distribution of evaluation items among participants is not entirely clear: While the paper specifies four participants, the published raw results only include a single quality assessment per item. In our reproduction, all participants evaluate all 63 test items, i.e. we collect 4 rankings per item and report the mean.

4 Reproduction Results

The results of the original study and our reproduction can be seen in Table 2. Per-evaluator results and significance levels are shown in Table 3.

System	Original	Reproduction	CV*
MemSum	**1.38**	1.49	25.21
NeuSum	1.57	**1.46**	21.3

Table 2: Original and reproduced scores (lower is better) and coefficient of variation (CV*, Belz 2022).

General Results With regard to the average ratings per system, our results differ notably from the original evaluation. In Gu et al. (2022), the proposed MemSum system achieves higher overall quality scores than the NeuSum method used as baseline. By contrast, NeuSum is slightly favored in our evaluation, although the average ranks diverge only marginally from a score of 1.5, which would indicate equal preference for both systems. Therefore, we were unable to confirm the main finding in Gu et al. (2022) that MemSum generates summaries of higher overall quality than the NeuSum baseline (cf. Section 2).

Coefficient of Variation Following the Extended Quantified Reproducibility Assessment (QRA++) framework (Belz, 2025) for *Type I* results, we report the unbiased coefficient of variation (CV*, Belz 2022) between the originally published results and the scores in our evaluation.[3] We rely on the implementation in Belz (2022), which is adjusted for small sample sizes. Since CV* requires metric scales to start at 0, but the quality scores in our evaluation are in a value range between 1 and 2, we we offset our results by -1 before calculating the CV*.

In line with our inability to reproduce the results of the original paper, the CV* scores are relatively high in our reproduction study (25.21 for MemSum and 21.3 for NeuSum, see Table 2).

Inter-Annotator Agreement We calculate the inter-annotator agreement between our evaluators using Fleiss's κ (Fleiss, 1971). Here, a score of $\kappa = 0.17$ only indicates *slight agreement* (Landis and Koch, 1977), pointing to notable differences between the ratings of the individual evaluators.

Per-Evaluator Results Table 3 shows the mean system ranks for individual evaluators. As an alternative, more interpretable measure, we also report the percentage of cases in which MemSum is rated higher than NeuSum. While three of the four eval-

[3] Assessments of *Type II* and *Type III* results are not applicable to this reproduction.

evaluator	MemSum	NeuSum	% MemSum #1	statistic	p
1	**1.44**	1.51	53.33	854.0	0.61
2	1.52	**1.43**	45.0	823.5	0.44
3	1.49	**1.46**	48.33	884.5	0.8
4	1.51	**1.44**	46.67	854.0	0.61
original	**1.38**	1.57	60.0	732.0	0.12

Table 3: Per-evaluator results and statistical significance tests (Wilcoxon signed-rank test, Woolson 2008) for ratings by individual evaluators and the ratings published in the original work. None of the rating series pass the significance threshold of $\alpha = 0.05$.

uators show a general preference for NeuSum, we note that all scores are close to perfect balance between the two systems (i.e., an average rank of 1.5 and a 50 % preference for MemSum), again pointing to weak overall tendencies.

Statistical Significance As in the original paper, we use a Wilcoxon signed-rank test (Woolson, 2008) to determine the statistical significance of the difference in ratings between MemSum and NeuSum. We apply this test to the ratings of all individual evaluators and to the ratings published by Gu et al. (2022). As shown in Table 3, none of the rating series pass the significance threshold of $\alpha = 0.05$. This includes all ratings of individual evaluators in our reproduction and the ratings originally published. However, we note that the p-value for the results in the original paper is considerably lower.

5 Discussion

As discussed in the previous section, we were unable to reproduce the main findings from Gu et al. (2022) with regard to the overall quality criterion in Experiment II. While the proposed MemSum model surpassed the NeuSum baseline in the original study, our results show the opposite trend, i.e., NeuSum is rated as better than MemSum on average. The high CV* scores corroborate these differences. However, it is important to note that the difference between the two systems is fairly small and not statistically significant, and the inter-annotator agreement reveals substantial differences in the judgments of individual evaluators. Reasons for the deviations from the original results and the measured uncertainty in our evaluators can be seen both in properties of the stimuli and in the experimental setup of this reproduction.

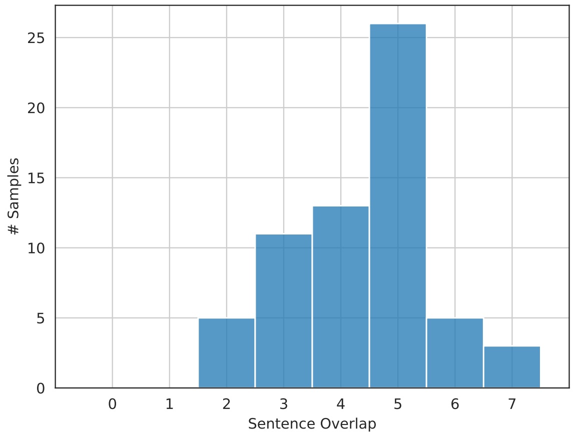


Figure 2: Sentence overlap between summaries extracted with MemSum and NeuSum. In more than 50 % of cases, generated summaries overlap by more than 5 sentences (with a total length of 7).

Properties of Stimuli As described in Section 3.1, Gu et al. (2022) used samples from the PubMed dataset for their evaluation. Importantly, this dataset consists of domain-specific texts from the medical field written in highly technical language, making it possible that evaluators, who are not domain experts, struggle to evaluate the content and textual quality of the summaries.

In addition, both systems often select similar sentences: As shown in the histogram in Figure 2, in more than half of the evaluation samples the outputs of the two systems overlap by at least 5 sentences, with a total length of 7 (see section 3.1). As a result, the summaries of both systems often only vary in detail, making it difficult to rank the methods.

Both of these aspects – the technical jargon and the high similarity between generated summaries – were named by participants as complicating factors subsequent to the evaluation.

Differences in Experimental Setups Another reason for the discrepancies between the original results and our reproduction may lie in the definition of the quality criterion. The notion of overall quality is relatively underspecified, which could lead to uncertainty regarding the exact properties of the texts against which they should be evaluated, although the evaluation interface provides somewhat more precise instructions, see Section 3.3. Importantly, as noted in Section 3.4, Gu et al. (2022) also included ratings for coverage and non-redundancy, which could affect the evaluation of overall quality — for example, if the ratings for these more specific criteria are included in the evaluation of overall quality.

Finally, as described in Section 3.4, the evaluation interface in the original study included an option to skip items if the summaries were too similar or if a decision could not be made for other reasons. Our interface lacks this function, forcing evaluators to decide on a ranking in all cases. Given the high similarity between the generated summaries for many items, this could be a reason for a higher rate of arbitrary decisions compared to Gu et al. (2022)'s evaluation, although it is unknown how many items were actually skipped in the original study.

6 Conclusion

In this paper, we attempted to reproduce the human evaluation from the work of Gu et al. (2022). Our evaluation produced clear differences from the original results, in particular we could not demonstrate that the proposed MemSum system produces summaries with higher overall quality than the baseline NeuSum system. At the same time, the narrow margin between the systems and the low inter-annotator agreement suggest fundamental uncertainties among our evaluators. To explain these discrepancies, we discussed differences between the original study and our reproduction, as well as general characteristics of the chosen domain and the generated summaries.

The mixed results in our study underline the problem that it is often difficult to reproduce the results of human evaluations in published papers, and stress the importance of projects like ReproHum.

Acknowledgments

We thank Simon Mille, Michela Lorandi and Rudali Huidrom for sharing their adapted version of the evaluation notebook, which was used for this reproduction.

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project no. 512393437, project B02.

References

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. 2025. Qra++: Quantified reproducibility assessment for common types of results in natural language processing. *Preprint*, arXiv:2505.17043.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

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

Anya Belz and Craig Thomson. 2024. Heds 3.0: The human evaluation data sheet version 3.0.

Anya Belz, Craig Thomson, Javier González-Corbelle, and Malo Ruelle. 2025. The 2025 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

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, and 23 others. 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.

Yanran Chen, Jonas Belouadi, and Steffen Eger. 2022. Reproducibility issues for BERT-based evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2965–2989, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Nianlong Gu, Elliott Ash, and Richard Hahnloser. 2022. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6507–6522, Dublin, Ireland. 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.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

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.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys*, 55(2):1–39.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. 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.

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.

R. F. Woolson. 2008. Wilcoxon signed-rank test.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.