ReproHum #0729-04: Partial reproduction of the human evaluation of the MemSum and NeuSum summarisation systems

Simon Mille and Michela Lorandi
ADAPT - Dublin City University, Dublin, Ireland
firstname.lastname@adaptcentre.ie

Abstract

In this paper, we present our reproduction of part of the human evaluation originally carried out by Gu et al. (2022), as part of Track B of ReproNLP 2025. Four human annotators were asked to rank two candidate summaries according to their overall quality, given a reference summary shown alongside the two candidate summaries at evaluation time. We describe the original experiment and provide details about the steps we followed to carry out the reproduction experiment, including the implementation of some missing pieces of code. Our results, in particular the high coefficients of variation and low inter-annotator agreement, suggest a low level of reproducibility in the original experiment despite identical pairwise ranks. However, given the very small sample size (two systems, one rating), we remain cautious about drawing definitive conclusions.

1 Introduction

In recent years, several editions of the ReproGen and ReproNLP shared tasks have been carried out –see, e.g., (Belz and Thomson, 2024a)-, which contributed to making the NLP community more aware of the importance of reproducibility when running and reporting on experiments. This year, the ReproNLP organisers proposed two tracks (Belz et al., 2025): Track A (*Open*) was for reproductions of any evaluation result, while for Track B (*ReproHum*), a set of 20 papers was preselected based on their suitability for being reproduced (availability of code, of instructions to evaluators, of detailed evaluation results, etc.). The present paper reports on one of the two reproductions for paper #0729-04 from Gu et al. (2022): *MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes*. In the following sections, we detail the original and reproduced experiments, the steps we had to take to run the evaluation, and the results of the reproduction study, discussing challenges encountered during the process.

2 Original experiment

This section contains a summary of the original experiment and a detailed description of the human evaluation procedure.

2.1 General experiment in original paper

In their paper, Gu et al. (2022) present the Multi-step Episodic Markov decision process extractive SUMmarizer (MemSum), which takes into account the extraction history when making decisions to extract a new span, so as to avoid redundancies and produce more compact summaries. They evaluate their system with ROUGE (Lin, 2004) on several English extractive summarisation datasets: PubMed and arXiv (Cohan et al., 2018), a truncated version of PubMed (Zhong et al., 2020), and GovReport (Huang et al., 2021). The authors show that MemSum obtains better metric evaluation than all baselines including state-of-the-art extractive and abstractive summarisers, i.e. NeuSum (Zhou et al., 2018) and Hepos (Huang et al., 2021) respectively.

2.2 Human evaluation in original paper

Gu et al. (2022) carry out two human evaluations that consist in ranking two summaries produced taking as input scientific articles from the PubMed data (Cohan et al., 2018):

- **Experiment 1** (67 pairs of summaries): [NeuSum summaries] VS [MemSum summaries with automatic stopping]; NeuSum summaries are always 7-sentence long, while MemSum summaries have no fixed length (5.6 sentences on average).
- **Experiment 2** (63 pairs of summaries): [NeuSum summaries] VS [MemSum summaries without automatic stopping]; both summaries contain exactly 7 sentences.

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 615–621

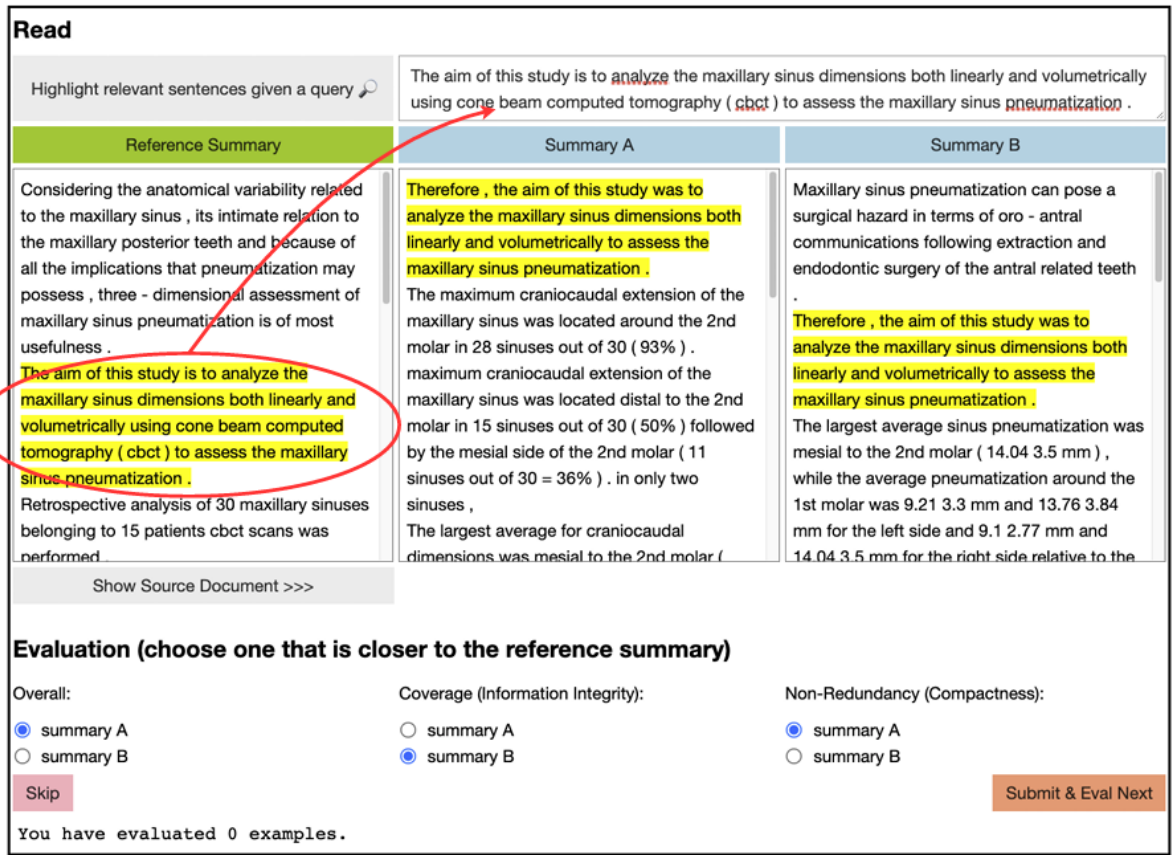

Figure 1: Original evaluation interface; copied from Appendix G in (Gu et al., 2022).

Quality criteria and evaluation operationalisation. In both experiments, four human evaluators assess three qualities of the summaries: *Coverage*, *Non-Redundancy*, and *Overall*. For each evaluation item, an evaluator sees three summaries: one reference summary on the left (from the PubMed dataset), then Summary A and Summary B (MemSum and NeuSum are randomly assigned A or B for each evaluation item). For each evaluation criterion, they have to choose which of Summary A or Summary B is "closer to the reference summary". *Coverage* is defined as "Information integrity" and *Non-Redundancy* as "Compactness", while *Overall* is not further specified.

User interface. The authors made available a user-friendly interface as a Google Colab Notebook; evaluators see the three summaries and the description of the criteria below them, along with a selection button to choose between Summary A and Summary B for each criterion. The interface also contains a highlighting tool: when participants type or paste spans of text into the box above the summaries, the text spans with a similar meaning are highlighted across all three summaries (see Section 3.3 for details on the implementation). The source documents from which the summaries were produced can also be shown/hidden. When the best system is selected for all three criteria, evaluators can submit the rankings and move to the next evaluation item. Figure 1 shows a screenshot of the original interface.

Computing results. For each criterion, the preferred system gets a score of 1, while the other system gets a score of 2. For each evaluation item, four scores are collected (one per evaluator). It is not entirely clear in the paper if the scores of the four annotators were aggregated at the item-level (via majority voting), and then averaged for each system (in this case, averaging 67 and 63 scores in Experiments 1 and 2), or if the scores of all evaluators were averaged for each system (in this case, averaging 67*4=268 scores in Experiment 1, and 63*4=252 scores in Experiment 2).

2.3 Additional information obtained from authors

The ReproNLP organisers contacted the authors to get additional information that was not clear in the paper. The authors confirmed that 4 evaluators took part to both experiments, and that all of them were

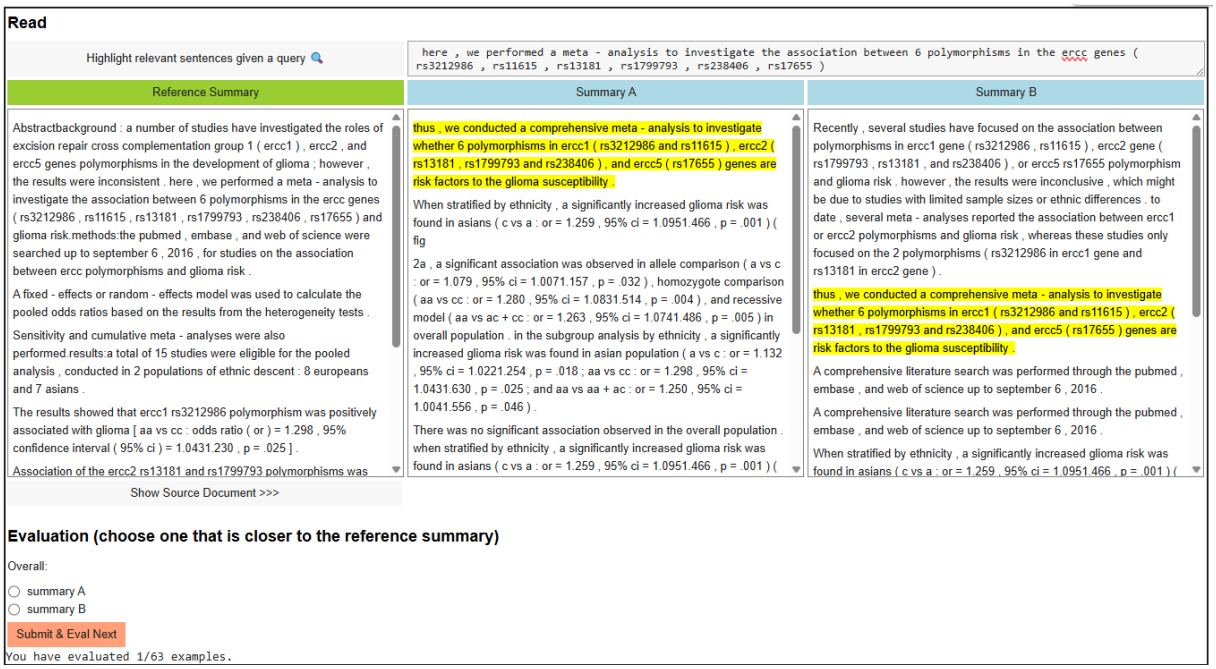

Figure 2: Evaluation interface for our reproduction study.

computer science students (PhD or Masters). The authors confirmed that they did not have another version of the Notebook than the one provided, in which some functionalities were missing (see Sections 3.3 and 3.4).

3 Our reproduction

In this section, we describe which experiment we reproduced and how we carried it out. All our code and documentation can be found on GitHub,[1] and details of our evaluation can be found in the Human Evaluation Data Sheet (HEDS) (Shimorina and Belz, 2022; Belz and Thomson, 2024b).[2]

3.1 The reproduced experiment

As specified by the ReproHum protocol, we carried out a reproduction of the evaluation of one criterion in one experiment, namely the **Overall** criterion of **Experiment 2** (see Section 2.2):

- **Experiment 2** (63 pairs of summaries): [NeuSum summaries] VS [MemSum summaries without automatic stopping]; both summaries contain exactly 7 sentences.

3.2 Evaluator recruitment and payment

As in the original study, we recruited four Computer Science Masters and PhD students as evaluators. Once the Ethics approval was obtained from the DCU Faculty Ethics committee, we sent an email to the NLP Masters and PhD students, and selected the first four students who answered. Our evaluators were either native English speakers or had English as a second language in which they are highly proficient. All evaluators read the experiment information sheet and then signed and returned the informed consent form before starting the evaluation. The task took them between 2 and 3 hours as planned, and each evaluator received a 50€ voucher as compensation for their time.

3.3 User interface

We were able to reuse the original experiment's Notebook, but some functionalities were missing so we had to (re)implement the following (see our interface in Figure 2):

- Highlighting functionality: as described in Section 2.2, the interface allowed for highlighting meaning-similar spans in the different summaries, but we could not find any function in the code which was triggered by entering text in the input field. Consequently, we reimplemented the highlighting function following the authors' description. Specifically, we used `sent2vec` (Pagliardini et al., 2018) to compute sentence embeddings for each sentence in Summary A and Summary B. Semantic similarity between sentences was

[1] https://github.com/mille-s/ReproHum_072904_DCU25

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

then assessed via cosine similarity. Sentences were highlighted if their similarity exceeded a predefined threshold t ($t = 0.6$). We used the same pre-trained embedding model used in the original study, i.e. the Wiki Unigram model.[3]

- Saving files: the Notebook we were provided was not saving the annotations. We added code to save the annotations in a Python pickle file every time the *Submit & Eval next* button was clicked. The pickle file was saved in the Google drive that was shared with the evaluators, which had two advantages: (i) every time the file was saved a new version of the file was created, which allows or recovering annotations in case something goes wrong; and (ii) partially completed files could be loaded, so that if the Notebook's runtime disconnected for some reason, the annotators could pick up where they left off. We implemented the loading functionality and integrated it in the Notebook.

- Cleaning of input json file: the provided files with the summaries to annotate already contained some scores from the original study; thus, we created a new json file in which we removed the scores so as to avoid any problem or ambiguity in the collected data.

In the shared drive, we created one notebook per evaluator; evaluators were assigned to a notebook via a shared spreadsheet.

3.4 Computing the results

No code was provided to compute the scores reported in the original paper, so we made our own version and added it to the Notebook. We implemented a simple function to load the newly connected annotations in Pandas data frames, from which we computed (i) the mean scores for each annotator for each of the two systems (mean of 63 scores for each system for each annotator, shown in Table 2), (ii) the mean score for each system across all four annotators (mean of 252 scores for each system, shown in the last column of Table 2 and at the bottom of Table 1), and (iii) the mean score for each system after aggregating the scores for each evaluation item (mean of 63 aggregated scores, shown at the top of Table 1). In the case of (iii), for each evaluation item we assigned 1 to the system that had the lower sum of scores across the four evaluators, 2 to the other system , and 1 to both systems in case of tie.[4]

While we assumed that calculating the mean score over all individual 252 scores for each system was the most natural way for computing the results, the results file found in the original repository contains only one score per evaluation item (63 scores), and when calculating the mean of these 63 scores for each system, we obtained the scores reported in the original paper (1.38 and 1.57 for MemSum and NeuSum respectively). We thus concluded that the authors aggregated the scores of the four evaluators for each evaluation item before computing the mean scores they reported, although we cannot exclude that the results correspond to one evaluator only, and that the mean scores of this evaluator are the same as the mean scores across all four evaluators. In Section 4 below, we report our results using both ways of calculating the mean scores.

3.5 Release and anonymisation of the data

The GitHub repository linked at the beginning of this section contains all the code we used in our reproduction, along with the anonymised evaluations collected in the process. In order for other teams to be able to carry out the same reproduction as we did, we also release a short guide for using the whole repository.

3.6 Known and possible deviations from original experiment

Several aspects of the method are not exactly as in the original experiment; we list them below as they could potentially have an impact on the results of this or future reproduction studies.

Number of criteria evaluated. The evaluators in our reproduction were not evaluating all three aspects but only one, which could have influenced their ratings.

Documentation. Since we modified the Notebook, we wanted to make sure that its functionalities were clear to the evaluators. We thus drafted some detailed instructions for using the Notebook and asked the participants to read them carefully before starting. Note that our instructions are limited to

[3] https://github.com/epfml/sent2vec?tab=readme-ov-file#downloading-sent2vec-pre-trained-models

[4] These are the three configurations we found in the original results file.

System	Original Study	Reproduction Study	Type I	Type II		Type IV
	Aggregated per item (?)	Aggregated per item	CV*	r	ρ	P
MemSum	1.38	1.27	33.74	-	-	1/1
NeuSum	1.57	1.33	53.17	-	-	1/1
	Aggregated per item (?)	Non-aggregated				
MemSum	1.38	1.47	21.11	-	-	1/1
NeuSum	1.57	1.53	7.25	-	-	1/1

Table 1: Comparison of original and reproduction mean scores for Gu et al. (2022)'s Experiment II's Overall criterion (we reproduced the original study scores with our code). Aggregated per item = mean score over 63 scores (one aggregated score per evaluation item); Non-aggregated = mean score over 252 scores (four scores per evaluation item). In each study, none of the score differences are statistically significant.

System	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Mean
MemSum	1.46	1.48	1.40	1.54	1.47
NeuSum	1.54	1.52	1.60	1.46	1.53

Table 2: Individual mean scores per evaluators in the reproduction study; IAA: 0.023 (Fleiss's κ).

the use of the interface, to remain as close as possible to the original study; the instructions to the annotators can be found in our GitHub.

The Skip button. In Appendix G of Gu et al. (2022), it is mentioned that the interface contained a *Skip* button (see Figure 1), which was to be used "if [the evaluators] were not sure which summary was indeed better". We however did not find the implementation of this button, and in the evaluation interface, there were no explicit instructions to evaluators that they could use it in case they could not decide between two summaries. Ultimately, we do not know if the Skip button was in the original user interface, and if it was, whether instructions for its use were provided to the evaluators. We decided to not provide a Skip button in the reproduction, which means that there is a possible deviation with respect to the original experiment.

Evaluators. The only thing we know about the original evaluation is that the evaluators were Master's and PhD computer science students; there can be differences in terms of age, gender, language proficiency, etc. between our evaluators and the original ones.

4 Results and discussion

Table 1 shows the original and reproduction scores for each system, along with the Quantified Reproducibility Assessment (QRA++) (Belz, 2025), which consists of (i) CV*, the coefficient of variation adjusted for small sample size (Belz, 2022), (ii) Pearson's r (which captures linear relationships) and (iii) Spearman's ρ (which captures monotonic relationships). The QRA++ numbers were computed using the QRA++ code provided by the organisers.[5] As discussed in Section 3.4, we were unsure as to how the mean scores were calculated for each system so we report two sets of scores which yield different mean scores and QRA++ results.

Quantified Reproducibility Assessment. Using the item-level aggregated scores, as was likely done in the original study, the CV* numbers are quite high, indicating a high degree of variation in the global results: 33.74 for MemSum and 53.17 for NeuSum. Using the mean of all individual rankings, the CV* is similar for MemSum, at 21.11, and considerably lower for NeuSum, at 7.25. Although these numbers are quite diverse, three of the CV* are greater than 20, which is a rather high number given previous reproduction studies; none of the CV* is below 5, which is usually associated with a low degree of variation. There are only two systems and they are ranked the same in both the original experiment and the reproduction, thus the Type IV result, namely the "proportion of identical pairwise system ranks" P (Belz, 2025), is 1 out of 1. We do not report Pearson's and Spearman's rank correlations in Table 1 because they do not bring any additional information with respect to P (both Spearman's and Pearson's correlations are maximal, at 1).

[5] As required by the QRA++ specifications, we offset our mean scores by -1 so the rating scale starts at 0, setting the *INSTRUMENT_SCALE_STARTS_AT* parameter at 1; i.e. the scores used for the first row are 0.38 and 0.27, instead of 1.38 and 1.27.

These QRA++ results thus suggest a low degree of reproducibility, and this is confirmed by further analysis: whereas there was a clear difference between the MemSum and NeuSum scores in the original experiment (0.19 points), the scores are more similar in our reproduction (0.06 points difference). As in the original paper, we ran the Wilcoxon signed-rank test (Woolson, 2005), and found no statistical significance at p=0.05 between the differences in scores for the two systems, be it using all 252 individual rankings (p value of 0.31) or the 63 aggregated rankings (p value of 0.52). Note that in the original experiment, the authors already reported no statistical significance between their Overall scores (p value of 0.12[6]). Our results suggest that the overall output quality of the different systems is possibly closer than reported in the original study.[7] This is confirmed by the examination of the individual evaluations discussed below.

Individual evaluators results. With respect to our individual annotator rankings, shown in Table 2, two evaluators (#1 and #2) have very similar mean scores while the other two (#3 and #4) have more polarised, but opposite, mean scores, one of them being almost identical to the average of the original experiment. In other words, in terms of mean scores, there is an apparent low agreement between our evaluators. We calculated the inter-annotator agreement using Fleiss's κ and obtained a score of 0.023, which indicates a poor agreement; this would certainly contribute to a high degree of variation in the results if the experiment were to be reproduced in the future.[8] These results confirm that the outputs of the two systems could be of comparable quality according to the unique criterion assessed in the study (Overall quality).

5 Conclusions

We conducted a reproduction study of Gu et al.'s (2022) Overall quality human evaluation of two summarisation systems, MemSum and NeuSum. Even though the outcome of our study is at first sight in line with the original study's results, MemSum achieving a slightly higher Overall score that NeuSum with no statistically significant difference, both our Quantified Reproducibility Assessment results (high coefficients of variation) and our detailed analysis of the global and per-annotator scores (marginal Overall system scores difference and a very low inter-annotator agreement) suggest a low level of reproducibility of the original study.

Thus, our interpretation of the evaluation results differs slightly from that of the original study: based on our analysis, the two systems appear to be very similar in terms of quality. This similarity may be attributed to both MemSum and NeuSum being extractive summarisers, with a significant proportion of the sentences selected by each system overlapping, which could make judgments difficult for annotators (i.e. because it is a relative evaluation, ranking two similar things is hard). However, considering the very small sample size (two systems, one criterion), we remain cautious in our interpretation. More reproductions would be needed to draw more solid conclusions.

Finally, although the reproduction process was not entirely straightforward and required some effort (see Section 3), we found that the majority of the necessary materials were available, and the reproduction in general was feasible and relatively smooth.

[6]Obtained by running our test on the original results file.

[7]In the original study, it is mentioned in Section 5.4 that MemSum "achieved a better average overall quality".

[8]For instance, almost half of the evaluation items (25/63) give a tied in ranking, i.e. two evaluators preferred one system, while two other evaluators preferred the other one.

Acknowledgements

Mille's contribution was funded by the European Union under the Marie Skłodowska-Curie grant agreement No 101062572 (M-FleNS), by the ADAPT research centre via the ADAPT Funding call 2024, and by the Irish Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media via the eSTÓR project. Lorandi's work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. We thank the authors of the original study for all their efforts into making their work reproducible. We thank the DCU Faculty Research Ethics Committee and Rudali Huidrom for their feedback on the Ethics approval application, and Craig Thomson for the support during the reproduction.

References

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 and Craig Thomson. 2024a. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia. ELRA and ICCL.

Anya Belz and Craig Thomson. 2024b. Heds 3.0: The human evaluation data sheet version 3.0. *Preprint*, arXiv:2412.07940.

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)*.

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.

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.

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.

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.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the 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.

Robert F Woolson. 2005. Wilcoxon signed-rank test. *Encyclopedia of biostatistics*, 8.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

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.