ReproHum#0043:

Human Evaluation

Reproducing Language Model as an Annotator: Exploring Dialogue Summarization on AMI Dataset

Vivian Fresen, Mei-Shin Wu-Urbanek, Steffen Eger
Adesso SE/Crif GmbH, Independent Researcher, Natural Language Learning Group (NLLG)
University of Mannheim
vivian.fresen@adesso.de, wumeishin@gmail.com, steffen.eger@uni-mannheim.de

Abstract

This study, part of the ReproHum Belz and Thomson (2024) project, a collaborative effort among researchers to replicate and assess experiments published in the natural language processing (NLP) literature, replicates and evaluates "Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization" by Feng et al. (2021). Using DialoGPT, BART, and PGN models, we assess dialogue summarization's informativeness on a scale of 1 to 5. Surprisingly, our findings diverge from the original study, with different models producing the highest-rated summaries. This discrepancy suggests limitations in reproducing the original results and underscores the need for further investigation into dataset selection and model effectiveness.

Keywords: keyword1, keyword2, keyword3

1. Introduction

Reproducibility in natural language processing (NLP) is crucial for reliability, to ensure that independent researchers can arrive at the same conclusions by following the original report's documentation. In NLP, reproducilibity extends beyond model training parameters and may involve the entire evaluation process leading to reported results. While reproducibility has been studied in NLP e.g., for automatic metrics or models (Fokkens et al., 2013; Post, 2018; Chen et al., 2022), there is a scarcity of work addressing human evaluation.

Human evaluation is particularly important, however, as human annotations most often provide the ground-truth against which NLP models are compared.

The work reported in this paper forms part of the ReproHum[1] project, which focuses on enhancing the documentation of human evaluation properties and evaluating the consistency between results obtained in reproduction studies and those of the original research Belz et al. (2023); Belz and Thomson (2024).

Our focus paper is Feng et al. (2021). We followed the paper's guidelines to reproduce the automatic summarization outputs by using DialogGPT (Zhang et al., 2020). To do so, we leveraged four PhD students to assess the generated texts. Our goal was to assess whether we could reproduce the original results along specific selected dimensions.

[1] https://reprohum.github.io/

Our report is structured as follows: Section 2 presents the original study design, providing an overview of the paper's content. In Section 3, we detail the reproduction of the NLP evaluation, outlining the specifics of the evaluation process to be replicated. Section 4 presents and discusses the results of the reproduced evaluation in comparison to the original paper. Finally, Section 5 offers concluding remarks and outlines avenues for future research.

2. Original Study Design

The original study, conducted by Feng et al. (2021), investigates enhancements to automatic text summarization. The study employs DialogGPT as an unsupervised annotator, focusing on three annotation aspects: keyword extraction, redundancy detection, and topic segmentation in dialogues.

Using DialogGPT, the authors annotate the SAMSum dataset (Gliwa et al., 2019) and the AMI dataset (Carletta et al., 2006), both containing dialogues and corresponding summaries. Pre-trained sequence-to-sequence BART (Lewis et al., 2020) and non-pretrained PGN (See et al., 2017) models are then used to generate summaries for the datasets annotated with keyword extraction (D_{KE}), redundancy detection (D_{RD}), topic segmentation (D_{TS}), and all three annotations combined (D_{ALL}) on both SAMSum and AMI datasets. The resulting summaries are assessed both automatically and manually.

BART(D_{KE}) demonstrates superior performance in the SAMSum dataset to the baseline and

Proceedings of the HumEval @LREC-COLING 2024, pages 199–209
Turin, Italy. May 21st, 2024.

PGN models, emphasizing the importance of keyword retention for concise dialogues. Conversely, PGN(D_{RD}) exhibits significant improvements in the AMI dataset, highlighting the necessity of redundancy detection.

The study aims to investigate whether incorporating DialogGPT as a component in text summarization, specifically through keyword extraction, redundancy detection, and topic segmentation, enhances the efficacy and potential improvements in dialogue summarization. This is achieved by comparing its results against established models using BLEU and ROUGE metrics. The corresponding human evaluation process aimed to assess the informativeness, conciseness, and coverage of dialogue summaries. A total of 100 dialogues from SAMSum and 10 meetings from AMI, along with their corresponding generated summaries, were randomly sampled for evaluation. Four human evaluators were hired to rate each summary on a scale of 1 to 5 for each metric, with higher scores indicating better performance (Feng et al., 2021).

3. Reproduction Study Details

We aimed to replicate the original study as precisely as possible. We used a subset of AMI dataset consists of ten dialogues, which is the identical material in Feng et al. (2021)[2]. The AMI Meeting Corpus is a rich multi-modal dataset containing approximately 100 hours of meeting recordings. It comprises both scripted scenario-based meetings, simulating design team collaborations, and naturally occurring meetings across various domains. The dataset includes audio, video, and transcript data, making it suitable for research in speech recognition, natural language processing, and human-computer interaction (Carletta et al., 2006). On the other hand, the SAMSum Corpus is a dataset designed specifically for abstractive dialogue summarization. It consists of chat dialogues that have been manually annotated with abstractive summaries. The corpus serves as a benchmark for evaluating automated summarization models tailored to the unique challenges posed by dialogue data (Gliwa et al., 2019) . The SAMSum Corpus offers a high-quality resource for researchers to develop and refine techniques for generating concise and informative summaries from conversational exchanges.

In our reproduction study, we focused solely on the AMI dataset and the informativeness criterion. By concentrating solely on one criterion, the reproduction experiment is simplified and easier to follow. Moreover, evaluating only one criterion should enable human evaluators to better concentrate on the annotation task; including multiple dimensions might confuse the annotators and hinder their ability to distinguish between the various criteria. However, this approach may potentially lead to overlooking other important aspects of dialogue summarization, resulting in less comprehensive evaluation outcomes.

Additionally, we followed the authors' guidelines and annotation outputs to be evaluated using the original scripts, without altering the summaries for each model and corresponding dialogue.

3.1. Evaluators

For the human evaluation, we engaged four annotators: native Chinese PhD students with high proficiency in English, as in the original study. One annotator is a PhD student in NLP, while the other three are from the fields of Sociology and Social Change. Each annotator received generous compensation,[3] as stipulated in the guidelines, for approximately 1-2 hours of work. The time estimation was based on the scope of the annotation task, which adhered to ReproHum recommendations. We adhered to the standardized ReproHum procedure for determining fair remuneration.

3.2. Differences to original study

The original study does not specify which interface was used for the Human Evaluators. We distributed the annotations to the evaluators via Google Forms (see Fig. 2 in the appendix), following the requirements set by the ReproHum team. This ensured uniform conditions and consistent result outputs for all reproduction experiments during the final evaluations and analyses. However, the outputs of the models, along with error annotations, remained consistent with those used in the original experiment.

In the original study, 100 dialogues from SAMSum and 10 meetings from AMI, along with their respective generated summaries, were assessed for informativeness, conciseness, and coverage by each model.

Fig. 1 shows an example of an AMI meeting with one of its summaries, followed by the option to rate the informativeness of the summary generated by the model. We were given the instruction to focus on the 10 AMI meetings only when reproducing the human evaluation, potentially to reduce annotation costs.

The instructions, originally provided in Chinese, were included with minor modifications by the authors. In Appendix A, we list them in the way we

[2]The full AMI dataset is provided in the repository on GitHub https://github.com/xcfcode/PLM_annotator.

[3]We paid each of them 50 EUR as a flat compensation in the form of amazon giftcards.

presented them for the human evaluation experiment. This approach facilitated a streamlined setup and enhanced accessibility for the annotators and the final evaluation process.

This is to prevent any potential influence on the reported outcomes. All information and resources should be accessed from the common resources folder provided by the project team. Any inquiries should be directed only to the ReproHum project managers, who communicated with the authors of the work being reproduced on behalf of the project.

4. Reproduction Results

We present our human evaluation result together with the scores provided in Feng et al. (2021) in Table 1. The comprehensive human evaluation results of the original article can be found in Table 4 in Appendix B.

The methodology for calculating the informativeness scores in the final evaluation results was not specified in Feng et al. (2021). Therefore, we utilized Python and R to calculate the informativeness scores over 10 AMI dialogs using three different methods: mean, median, and mode. Additionally, we adopted Feng et al. (2021)'s approach of using Fleiss' kappa score for evaluating inter-annotator agreement in our study. The scripts to calculate the mean, median, mode, as well as Fleiss' kappa scores are provided in our github repository [4].

	Model	Original	Mean	Median	Mode
	Golden	4.70	2.4	2.5	3
AMI	PGN	2.92	2.18	2.0	2
	HMNet	**3.52**†	2.2	2.0	2
	PGN(D_{KE})	3.20	2.18	2.0	2
	PGN(D_{RD})	3.15	**3.0**††	**3.0**	**3**
	PGN(D_{TS})	3.05	2.27	2.0	1
	PGN(D_{ALL})	**3.33** ††	**2.52**†	**3.0**	**3**

Table 1: Human evaluation results from Feng et al. (2021) is provided in the 'Original'column. The informativeness result in the reproduction experiment is provided in the 'Mean', 'Median' and 'Mode'columns. The corresponding Fleiss' kappa scores in the original paper are 0.48. The Fleiss' kappa score of our reproduction experiment is 0.069.

Findings Comparison The original results presented in the paper by Feng et al. (2021) indicate that their method, which combines DialogGPT as an annotator with BART and PGN as summarization generators, achieved the highest scores. Particularly, the combination of DialogGPT Redundancy reduction (D_{RD}) with both BART and PGN resulted in better scores for conciseness (another dimension of annotation not considered by us). Additionally, when combined with DialogGPT Topic Segmentation (D_{TS}),the model performed better in coverage. However, HMNet, a Hierarchical Memory Network,[5] attained the best scores in informativeness and coverage for the AMI dataset.

There is a decisive gap between the scores of generated summaries and the scores of the gold summaries in the original study, indicating the increased difficulty of the AMI dataset (Feng et al., 2021). However, we did not observe such a significant difference between the score of Gold standard and the informativeness scores of the AMI dataset in our experiment. In Fig. 1 we can see an example of an AMI meeting provided for the human evaluation experiment, and its summary with the respectively rating options for informativeness. In Section C of the appendix, we give some examples where our raters disagree with the raters of the original study.

Our result in contrast with the original study is shown in Table 1.

The Informativeness measure applied to the gold outputs demonstrates a significant coefficient of variation(CV*) of 64.59%, indicating substantial variability relative to the mean value of 3.55. The unbiased sample standard deviation of 2.038 highlights considerable dispersion around the mean within the dataset. However, due to the small sample size of 2, the reliability of the standard deviation as a measure of dispersion may be limited.

Table 2 shows the coefficient of variation (CV*) with the corresponding mean values. The CV* metric is adapted for small sample sizes, making it suitable for use even with the limited pairs of results one may have (Belz, 2022).

Table 2: Coefficient of Variation (CV*) with Mean

Sample	Mean	CV*
1	3.55	64.59
2	3.22	18.58
3	3.47	15.52
4	3.10	3.22
5	2.19	1.14
6	2.59	31.79
7	2.40	10.41

Table 3: *
Note: CV* denotes the Coefficient of Variation.

[4] https://github.com/vivianCF/HumanEvaluation.git

[5] HMNet is a state-of-the-art model designed for abstractive dialogue summarization. It leverages memory modules and hierarchical attention mechanisms to capture dialogue nuances effectively. By storing relevant information and attending to different dialogue levels, HMNet generates coherent and informative summaries that faithfully represent the input dialogue.

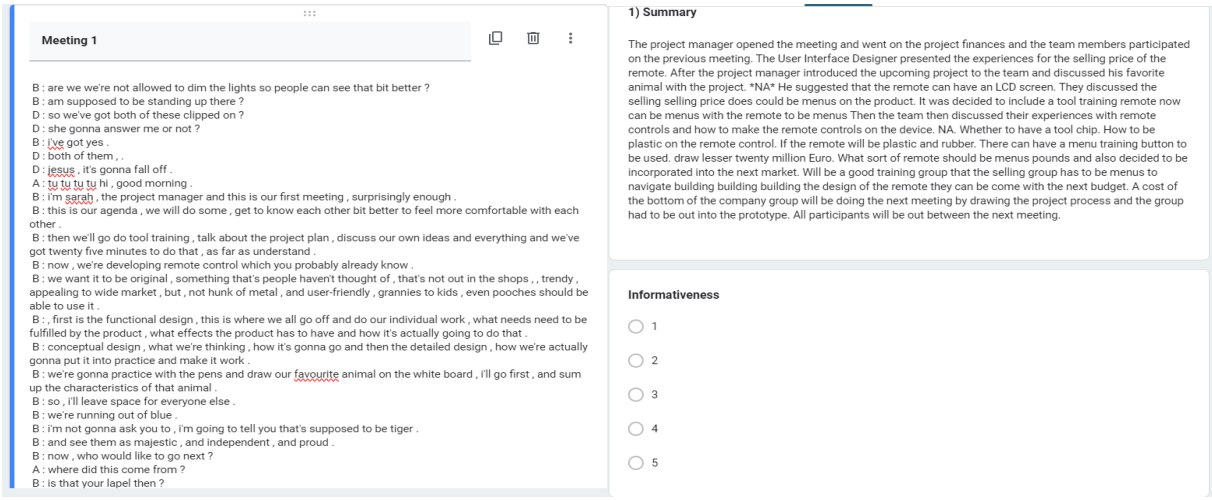

Figure 1: Example of an AMI meeting and its summary from one model.

Additionally, the wide confidence interval (-9.62 to 13.70) accentuates the uncertainty surrounding the true population mean, necessitating cautious interpretation of the dataset.

The coefficient of variation for PGN is 28.93%, indicating moderate variability relative to the mean value of 2.55. The sample standard deviation of 0.656 supports this observation, suggesting a moderate dispersion of data points around the mean. The confidence interval (-3.095, 4.407) implies some uncertainty about the true population mean. However, all measured values fall within one and two standard deviations from the mean, indicating a symmetric distribution around the mean.

HMNet exhibits a higher coefficient of variation at 46.02%, indicating high variability compared to PGN relative to the mean value of 2.86. The sample standard deviation of 1.170 suggests a greater dispersion of data points around the mean. The wider confidence interval (-5.521, 7.861) implies more uncertainty about the true population mean compared to Dataset 1. However, all measured values still fall within one and two standard deviations from the mean.

Similarly, PGN(D_{KE}) shows a coefficient of variation of 37.80%, indicating moderate variability compared to Dataset 1 relative to the mean value of 2.69. The sample standard deviation of 0.904 supports this, suggesting a moderate dispersion of data points around the mean. The confidence interval (-4.266, 6.074) also indicates some uncertainty about the true population mean. Nonetheless, like the other datasets, all measured values fall within one and two standard deviations from the mean.

PGN(D_{RD}) has the lowest coefficient of variation (4.86%), indicating the least variability compared to the mean value of 3.075 among all datasets. The sample standard deviation is also the smallest (0.133), suggesting minimal dispersion of data points around the mean.

In Contrast,PGN(D_{TS}) shows again a higher coefficient of variation (29.24%) compared to PGN(D_{RD}), indicating higher variability relative to the mean value of 2.66. The sample standard deviation is also larger (0.691), suggesting a greater dispersion of data points around the mean.

PGN(D_{ALL})shows a coefficient of variation of 27.61%, slightly lower than PGN(D_{TS}), indicating slightly less variability compared to the mean value of 2.925. The sample standard deviation (0.718) is comparable to PGN(D_{TS}), suggesting a similar dispersion of data points around the mean.

In summary, PGN(D_{RD}) demonstrates the least variability, followed by PGN(D_{ALL}) and PGN, respectively. Overall, despite variations in coefficient of variation and sample size, all datasets exhibit symmetric distributions around the mean, as indicated by all measured values falling within one and two standard deviations from the mean.

Both the original study and our reproduction experiment suggest that PGN combined with redundancy reduction can achieve good performance for the AMI dataset in dialogue summarization. However, the gap between the gold standard and the other datasets in our case is not substantial, with a score of 2.4; the score is still considerably lower than the original result of 4.70.

In summary, these findings indicate a significant deviation compared to the original study. There are no significant discrepancies observed between the gold standards and the remaining datasets in our experiment, suggesting a different behavior compared to the original study. We observe a distinct trend compared to the original study; for instance, in our experiments, PGN (D_{RD}) demonstrates the highest performance, with PGN (D_{ALL}) closely fol-

lowing, which is somewhat unexpected considering that in the original study, HMNet achieved the highest score followed by PGN (D_{ALL}).

In our reproduction experiment, the scores were overall inferior to those in the original study, mostly all below 3.0 versus the original scores were all above 3.0. Specifically, the gold standard scores in our analysis are significantly lower than those reported in the original study. In our experiment, we have noticed a distinct trend that contrasts with the findings of Feng et al. (2021) in which the performance of HMNet does not exhibit substantial gains over the PGN models.

Upon comparing the coefficient of variation (CV*), it becomes evident that more replications of the same experiment may be required to draw more robust conclusions about the results presented in the human evaluation reproduction approach.

Furthermore,in our annotation task, we did not achieve comparable agreement (0.069), for AMI and informativeness on the same dataset. This is in strong contrast to the original study, which reported agreements of above 0.40 throughout.

5. Conclusion

Our research focuses on the reproduction and evaluation of dialogue summarization models through human assessment. The collaboration with the ReproHum organizers and access to materials from the original authors greatly facilitated the successful execution of our reproduction experiment.

Our key findings include:

- In our reproduction study, the inter-annotator agreement was notably lower, registering at 0.069, compared to above 0.40 reported in the original study.

- We were unable to confirm the effectiveness of the proposed approach in terms of informativeness. While we observed a moderate positive Pearson correlation coefficient of 0.481 between the informativeness of the original study and our experiment, indicating a medium level of correlation, the Spearman correlation coefficient of approximately -0.058 between both experiments suggests a weak negative monotonic relationship. Overall, the correlations between the original human evaluation and our reproduction are weak. However, it must be kept in mind that we correlated vectors of very short size (length seven).

  Notable discrepancies in human evaluation outcomes persist, indicating potential variations in annotators, methodology, or dataset selection for each dialogue summarization. We hypothesize that these differences could be attributed to two main factors. Firstly, the dataset is limited to only 10 meeting materials, which may lead to skewed average scores, favoring extreme values. Additionally, the involvement of only three evaluators may not provide a comprehensive assessment. Further experiments and reproductions are necessary to draw more conclusive findings from this study.

- The evaluated model performances in our reproduction study were inferior compared to the scores reported in the original study.

- The ratings in the “Original" column of Table 1 were not elucidated in the original study. From the context, we inferred the authors reported the average scores. However, in a small-scale study, one method to prevent outliers from impacting the mean is to utilize the median. Consequently, we were unsure whether it is indicative of a mean or a median.

- The human evaluation in our case is neither repeatable nor reproducible.

A potential explanation for these results is the persistent gap between the scores of generated summaries and those of gold summaries for the AMI dataset, indicating its inherent difficulty. The complexity and ambiguity of the dialogues posed a challenge during the experiment's preparation, making them difficult to follow and leading to divergent ratings among evaluators.

Moreover, the original study regarding AMI/Informativeness did not demonstrate effectiveness: the baseline HMNet performed the best. This raises the question of whether the selection of the AMI dataset is appropriate for the human evaluation reproduction and the verification of the performance of the models using DialoGPT to achieve better performance in dialogue summarization. Furthermore, conducting a comprehensive analysis of dataset characteristics and evaluation metrics could offer valuable insights into enhancing the appropriateness of the dataset selection for evaluating summarization models.

Our reproduction study raises an intriguing question about the identification and management of subjective practices that might have been employed in the original study. The lack of information on human participants' training depth and the undisclosed time investment in annotations during the original study contribute to uncertainties in interpreting the significant disparity in our human evaluation results.

Acknowledgements

We thank the ReproHum Group for guidance on conducting the experiment and for providing information and feedback. Special thanks to CRIF GmbH for providing space and time, making the completion of this project possible. The NLLG group gratefully acknowledges support from the BMBF via the grant “Metrics4NLG” and the DFG via the grant EG375/5-1.

6. Bibliographical References

Anya Belz. 2022. A metrological perspective on reproducibility in nlp. *Computational Linguistics*, 48(4):1125–1135.

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. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A preannouncement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. 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.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

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.

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.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*,

pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

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.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What have we learned from the checklist? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12789–12811, Toronto, Canada. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. 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.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. 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.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. 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.

Appendix

A. Annotator Guidelines

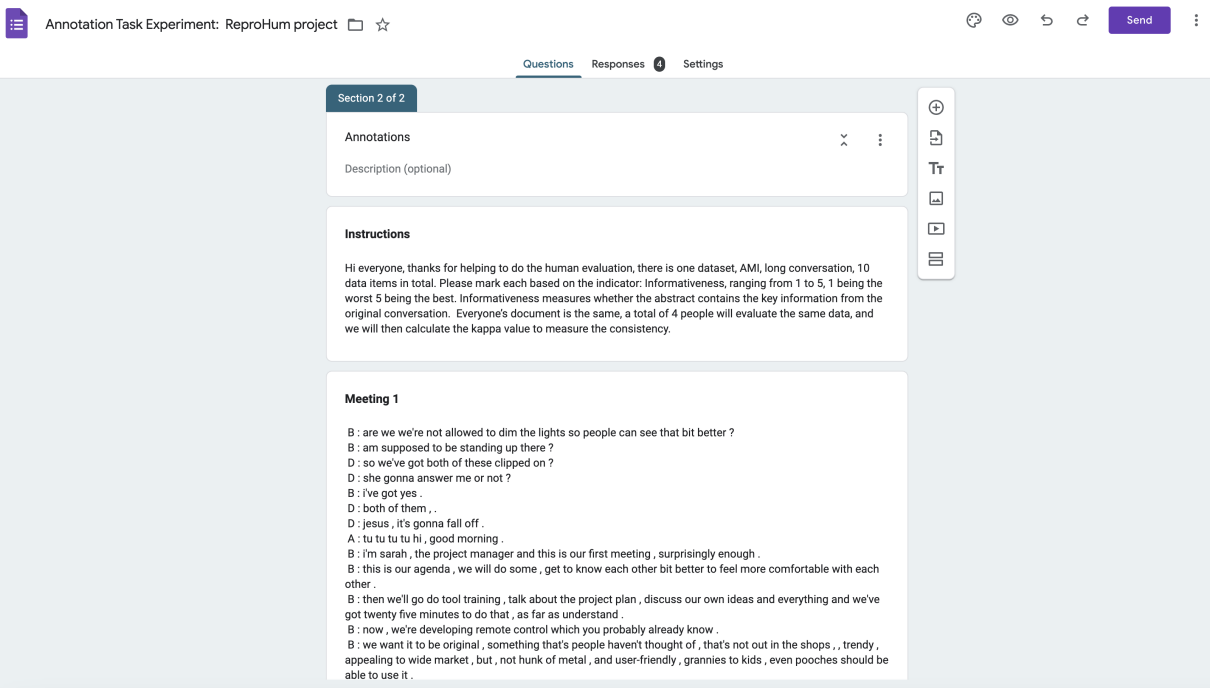

Figure 2: Example of Google Forms interface used during the Human Evaluation reproduction experiment

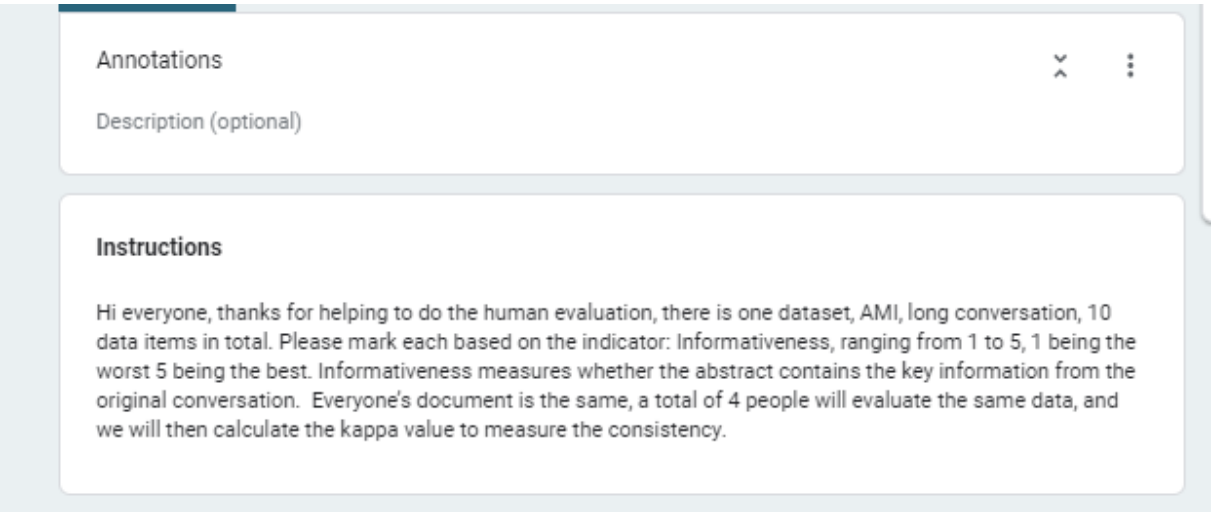

Figure 3: Example of AMI instructions for Human Evaluation

B. Results from original Study

	Model	Info.	Conc.	Cov.
	Golden	**4.37**	4.26	4.27
SamSum	BART	3.66	3.65	3.66
	MV-BART	3.85	3.76	3.88
	BART(D_{KE})	3.88	3.77	3.79
	BART(D_{RD})	3.74	**3.98**†	3.89
	BART(D_{TS})	**3.95**†	3.76	**4.01**††
	BART(D_{ALL})	**4.05**†	**3.78**††	**4.08**†
	Golden	4.70	3.85	4.35
AMI	PGN	2.92	3.08	2.70
	HMNet	**3.52**†	2.40	**3.40**†
	PGN(D_{KE})	3.20	3.08	3.00
	PGN(D_{RD})	3.15	**3.25**†	3.00
	PGN(D_{TS})	3.05	**3.10**††	**3.17**††
	PGN(D_{ALL})	**3.33**††	**3.25**†	3.10

Table 4: Human evaluation results from the original paper indicate the following abbreviations: 'Info.' for informativeness, 'Conc.' for conciseness, and 'Cov.' for coverage. In the case of SAMSum, the inter-annotator agreement (Fleiss' kappa) scores for each metric are 0.46, 0.37, and 0.43, respectively. For AMI, the corresponding Fleiss' kappa scores are 0.48, 0.40, and 0.41.

Fleiss' Kappa Value	Interpretation
0.00 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost perfect agreement

Table 5: Interpretation of Fleiss' Kappa Values

C. Summary Examples

2) Summary

the project manager introduced the project to the team and then the team members participated in an exercise in which they drew their favorite animal on the white board and discussed why they liked the particular animal . the project manager discussed the project finances and selling prices . the group then evaluated the project process , and discussed their experiences with remote controls . they discussed making the remote universally compatible , to be a mobile phone , and using plastic instead of metal . the user interface designer discussed the controls and how they operated together , and presented the type of scroll in the shared design . the marketing expert led an evaluation of the prototype . the remote will be made of plastic and will feature chunky colors and voice . production costs cannot exceed 12 .50 euro . monkey control will be small , appealing to a wide market , but not a metal user youth . having a feature on the remote which allows the user to locate a lost seagull . having discussion about a dual function on the screen , channel up , two basic functions , and one for the basic functions of the user interface . she shares the list of the features that will be incorporated into the design of a remote that should be used . she will run out of the blue . the sheepdog will be a flying seagull . the marketing expert will work on trend watching the remote control will not contain a useful feature . will have a menu , display , menu , channel , and menu buttons . would like to see if it is misplaced by someone like it , so that the eagle was not a vampire bat and will have to be used with plastic . she had to use plastic for the display since the remote controls were too expensive . however , the group decided that they had to re it every time and to eliminate the signals . it might not be too costly to add to look good and incorporate a feature that has since the product is unlikely that the remote will make it desirable and be fashionable so that it can be a soft and stylish , fashionable and cool and cool , traditional remote .

Figure 4: Example of summary (HMNet) for the meeting 1 of AMI data set.

3) Summary

The project manager opened the meeting and introduced the upcoming project to the team members introduce themselves and then decided over the agenda. The team then discussed their experiences from the previous meeting. They discussed the project process and discussed the cost of the remote. *NA* *NA*. It will be made of rechargeable with a docking station and sleek theme. He discussed the extra buttons that should be incorporated into the design. After the project manager talked about the project finances and the team discussed the features they could be made from the next meeting and has them as them as they liked about the cost for the remote controls and they decided to incorporate a docking station. Then the team evaluated the project budget and the group had to focus with the project. Overall, the group decided to include a wide with the remote control. NA. Whether the user interface designer will look at the functional requirements and has a remote controls was hard buttons could be incorporated to the budget of the product. How to use a plastic which should be used. What sort of extra buttons or the remote will be made Will not have a wide station that should have a trendy station as they could have a teletext function. There will not have an LCD screen. Size of rechargeable screen and corporate image that they want to use up to the project budget. A extra buttons are more

Figure 5: Example of summary (PGN(D_{KE})) for the meeting1 of AMI dataset.

4) Summary

The project manager opened the meeting by stating the agenda. The industrial designer discussed the interior workings of a remote which is easier to be He discussed the target group of the features of the function of the device. They also discussed using a timer be an LCD screen. After the project manager closes the meeting and going the team members introduce themselves by name and the team discussed their favorite animal and discussed what features they wanted to include for the remote they be easier to use. Then the team found the project finances and what features the target goals. *NA* NA. and at the functional functional group to be out from the working design. It was decided to include a timer to use a timer which can only be used for the television or the television remote which will only be set on the functional design, and the the marketing will be no limited at a regular target group to have a feature to address their product. There will have a hook screen which is no at address a feature of the remote. A decision that the remote would have a display function. Whether to have an LCD screen which will be used to address the and the television target group had to be no limited for the budget. When they have not sure their project process and the project manager's closing, the User Interface Designer and the Industrial Designer to research possible important and the group decided to use an LCD screen, and the Marketing Expert to prepare the working design of the remote and that they will include to include the remote which should be easier

Figure 6: Example of summary (PGN(D_{TS})) for the meeting10 of AMI dataset.

1) Summary

The project manager opened the meeting and went on the project finances and the team members participated on the previous meeting. The User Interface Designer presented the experiences for the selling price of the remote. After the project manager introduced the upcoming project to the team and discussed his favorite animal with the project. *NA* He suggested that the remote can have an LCD screen. They discussed the selling selling price does could be menus on the product. It was decided to include a tool training remote now can be menus with the remote to be menus Then the team then discussed their experiences with remote controls and how to make the remote controls on the device. NA. Whether to have a tool chip. How to be plastic on the remote control. If the remote will be plastic and rubber. There can have a menu training button to be used. draw lesser twenty million Euro. What sort of remote should be menus pounds and also decided to be incorporated into the next market. Will be a good training group that the selling group has to be menus to navigate building building building the design of the remote they can be come with the next budget. A cost of the bottom of the company group will be doing the next meeting by drawing the project process and the group had to be out into the prototype. All participants will be out between the next meeting.

Figure 7: Example of summary (PGN(D_{RD})) for the meeting10 of AMI dataset.

D. HEDS

Below is the HEDS of the Human Evaluation Experiment. The original HEDS are based on the provided documents of the ReproHuman Group (Shimorina and Belz, 2022), which can be accessed at: https://favorite-fox.static.domains/heds-2022-11-18. For more information and updates, please visit the ReproNLP 2024 GitHub repository: https://github.com/nlp-heds/repronlp2024.