ReproHum #0067-01: A Reproduction of the Evaluation of Cross-Lingual Summarization

Supryadi, Chuang Liu, Deyi Xiong*
TJUNLP Lab, College of Intelligence and Computing, Tianjin University, Tianjin, China
{supryadi, liuc_09,dyxiong}@tju.edu.cn

Abstract

Human evaluation is crucial as it offers a nuanced understanding that automated metrics often miss. By reproducing human evaluation, we can gain a better understanding of the original results. This paper is part of the ReproHum project, where our goal is to reproduce human evaluations from previous studies. We report the reproduction results of the human evaluation of cross-lingual summarization conducted by Bai et al. (2021). By comparing the original and reproduction studies, we find that our overall evaluation findings are largely consistent with those of the previous study. However, there are notable differences in evaluation scores between the two studies for certain model outputs. These discrepancies highlight the importance of carefully selecting evaluation methodologies and human annotators.

1 Introduction

In recent years, natural language processing (NLP) has witnessed remarkable progress, driven by advances in NLP models and data sources. This progress has led to significant improvements across a wide range of NLP tasks, including machine translation (Supryadi et al., 2024), text summarization (Hasan et al., 2021), reasoning (Shi et al., 2024b), and question answering (Yu et al., 2024). Evaluation plays a crucial role in assessing NLP models before they are deployed in real-world applications (Guo et al., 2023; Shi et al., 2024a). NLP model evaluation is typically conducted using automated metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). In addition to these metrics, human evaluation also plays an important role by providing insights into model performance based on human preferences and real-world applicability.

Reproduction studies are crucial for ensuring the reliability and quality of research experiments, especially for human evaluation. They help verify the validity of findings and build trust in scientific results. However, reproduction can be challenging due to missing information and lack of detailed documentation in previous experiments (Belz et al., 2023). The ReproHum project (Belz and Thomson, 2024) organises a shared task to investigate the extent to which human evaluation experiments are reproducible.

As part of the ReproHum project B batch experiment (Belz et al., 2025), we focus on reproducing the human evaluation conducted in the paper "Cross-Lingual Abstractive Summarization with Limited Parallel Resources" by Bai et al. (2021). The original study aims to improve cross-lingual summarization in low-resource settings. Specifically, for the human evaluation, they assessed 60 Chinese paragraphs with four different English summarization results each.

In this paper, we first detail the experiments conducted in the original research, with a specific focus on human evaluation in Section 2. We then introduce our reproduction setting in Section 3. Finally, we report the quantified reproducibility assessment and compare the results of our reproduction study with those of the original study in Section 4.

2 Original Study

The study we are focusing on reproducing is "Cross-Lingual Abstractive Summarization with Limited Parallel Resources" by Bai et al. (2021). In the original study, the authors proposed Multi-Task Cross-Lingual Abstractive Summarization (MCLAS), a framework designed to enhance cross-lingual summarization in low-resource settings. The model employs a pre-training and fine-tuning strategy. Initially, it is pre-trained on a large-scale monolingual document-summary dataset to equip the decoder with general summarization capabilities. Subsequently, it is fine-tuned on a small num-

*Corresponding author.

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 609–614

ber of parallel cross-lingual summary samples to transfer the learned summarization capabilities to low-resource languages.

2.1 Dataset and Models

The datasets used in the experiments include Zh2EnSum (Chinese-to-English) and En2ZhSum (English-to-Chinese) (Zhu et al., 2019). Additionally, a new En2DeSum (English-to-German) dataset was constructed. These datasets vary in size and are used to evaluate the model's performance in both low-resource scenarios (with minimum, medium, and maximum sample sizes) and full-dataset scenarios of training samples for all datasets. For the baselines, the authors compared neural cross-lingual summarization (NCLS) and neural cross-lingual summarization + monolingual summarization (NCLS+MS) (Zhu et al., 2019).

2.2 Human Evaluation

They also conducted human evaluations to examine the model performance. First, they randomly selected 60 examples (20 for each low-resource scenario) from the Zh2EnSum test dataset. Seven graduate students proficient in English and Chinese evaluated three generated summaries (MCLAS, NCLS, NCLS+MS) and gold summaries, focusing on informativeness (IF), fluency (FL), and conciseness (CC). IF assesses the importance of the extracted information, CC evaluates whether the summary is concise and free of redundant information, and FL checks the grammar and syntax fluency of the summaries.

The evaluation used the Best-Worst Scaling method (Kiritchenko and Mohammad, 2017), where participants chose the best and worst items for each perspective. Final scores were calculated based on the percentage of times each system was selected as best minus the times it was selected as worst, ranging from -1 (worst) to 1 (best). The results showed that MCLAS outperformed NCLS and NCLS+MS in all metrics, particularly in conciseness. The Fleiss' Kappa scores and overall agreement percentages indicated good inter-observer agreement among participants.

3 Reproduction Settings

In this study, we focus on reproducing the human evaluation from the original study. We express our gratitude to the original authors for sharing the experiment data, from the evaluation forms and the anonymized annotation results. With this data, we can compare our reproduction results with the original study.

We filled Human Evaluation Datasheet (HEDS), a document containing the comprehensive details for the human evaluation reproduction experiment. The HEDS document is available in a GitHub central repository.[1]

3.1 Human Annotators and Annotation Platform

We followed the annotator requirements outlined by the original authors by recruiting 7 students proficient in both English and Chinese. Upon further inquiry with the authors, we learned that these students were master's students and labmates of the authors, actively engaged in NLP research. Similarly, we recruited 7 master's students from our university's NLP laboratory, to ensure consistency in the evaluation process.

The previous author reported that the annotation platform is currently inaccessible. Therefore, we use another platform for the annotation. We considered using WeSurvey,[2] an open-source questionnaire platform by Tencent in China. We chose this platform because the participants are based in China, and it offers greater accessibility and convenience.

3.2 Evaluation Annotation Design

We conducted the experiments by distributing a questionnaire link to respondents. Upon opening the link, respondents see a consent form. This form confirms that the research has been explained, they can ask questions, and their anonymized data will be used for research purposes. They can withdraw at any time before data anonymization. If they agree, they proceed to complete the questionnaire.

Next, we collect the respondents' names and email addresses to send them vouchers upon completing the questionnaire. We also inquire about each respondent's English language proficiency. Additionally, we verify that the respondents are indeed master students studying NLP.

We follow the previous study by using 60 examples, with 20 examples for each of the three different low-resource scenarios (minimum, medium, and maximum). However, this study differs in its focus, as it evaluates only the "informativeness"

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

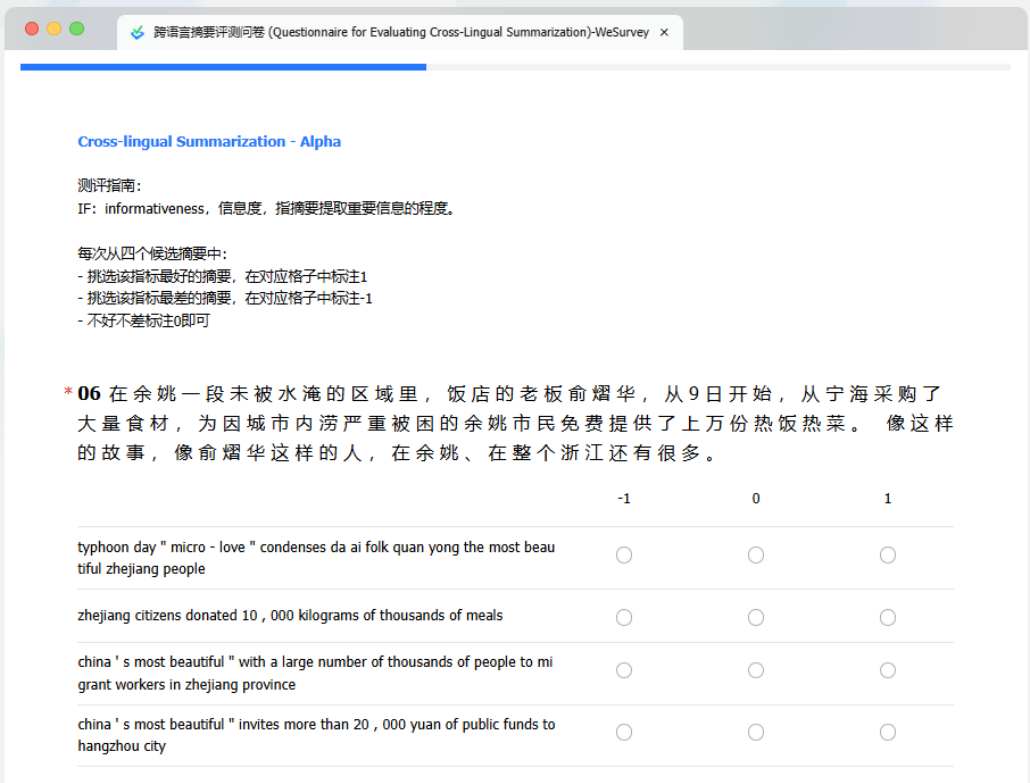

Figure 1: Screenshot of the annotation platform used in reproduction study.

metric, without including "conciseness" or "fluency". For each set of four candidate abstracts, participants need to select the summarization result with the highest informativeness and mark "1" in the corresponding grid. The result with the lowest informativeness need to be marked "-1" in the matching grid. All remaining grids will be filled with "0".

The screenshot of our questionnaire is shown in Figure 1. We label each scenario as Alpha, Beta, and Delta. Given that our respondents are Chinese students, the instructions are in Chinese. We explain the "informativeness" metric, which measures the important information extracted in the summary. For annotation, respondents are instructed to mark the best summarization result as 1, the worst as -1, and the others as 0. The example includes a Chinese paragraph with four English summarization results.

3.3 Payment

We follow the approach of previous studies by compensating participants for evaluating the summarization results. Specifically, we provide JD vouchers, a shopping voucher in China valued at approximately 100 RMB, as a token of appreciation for their participation in annotating the datasets.

Scenarios	Models	Original	Repro	CV*
Minimum	MCLAS	-0.264	-0.329	9.21
	NCLS	**-0.243**	**-0.093**	17.97
	NCLS+MS	-0.371	-0.264	15.63
	GOLD	0.879	0.686	10.8
Medium	MCLAS	0	**-0.007**	0.7
	NCLS	**0.036**	-0.214	27.36
	NCLS+MS	-0.343	-0.3	6.32
	GOLD	0.3	0.521	15.62
Maximum	MCLAS	**0.057**	**0.079**	2.05
	NCLS	-0.129	-0.129	0
	NCLS+MS	-0.179	-0.193	1.71
	GOLD	0.257	0.25	0.56

Table 1: Human evaluation results from original paper and reproduction experiment for "informativeness" metric. We also present the CV* score. The best score is bolded without comparing the gold summarization results.

Scenario	r	r (p-value)	ρ	ρ (p-value)
Minimum	0.98	0.019	0.8	0.2
Medium	0.86	0.138	0.8	0.2
Maximum	0.99	0.003	1	0

Table 2: The pearson (r) and spearman (ρ) correlation between original and reproduction study for each different scenarios.

15. 上半年被业界津津乐道，甚至被当成是推动国内电信业改革的号角的虚拟运营商，却于前几日被曝出了令人大跌眼镜的成绩单。自五月开启放号，十几家虚拟运营商总共放出仅约20万个号码，而实际活跃用户更是只有2万人左右。

Translation: The virtual operators, which were talked about by the industry in the first half of the year and even regarded as the clarion call for the reform of the domestic telecommunications industry, were exposed to shocking performance a few days ago. Since the number release began in May, more than a dozen virtual operators have released only about 200,000 numbers in total, and the actual number of active users is only about 20,000.

virtual operator : a plate of fresh meat , broken in the pot	1
first half of spontaneously : more than 20 , 000 users	-1
170 , 000 mobile phone numbers have been blackmailed by two , and less than 2 , 000	-1
170 mobile operators were blackmailed in the first half of the year : less than 2 , 000	-1

Figure 2: Example of error annotation.

Models	Krippendorf's α
MCLAS	0.135
NCLS	0.130
NCLS+MS	0.160
GOLD	0.204

Table 3: Krippendorff's α results from original and reproduction study for each model.

Claim
Claim 1: As the data size increases, all the models achieve better results.
Claim 2: MCLAS outperformed NCLS and NCLS+MS in all the metrics
Claim 3: MCLAS is especially strong in conciseness.

Table 4: Claims from original experiment.

4 Quantified Reproducibility Assessment

The evaluation in this study follows the standardized procedure established by the ReproHum project, which categorizes reproducibility into four types of results (Belz, 2025). In Type I, we report the single numerical scores and coefficient of variation (CV) values. For Type II, we calculate both Pearson and Spearman correlation coefficients. In Type III, we present an agreement score that quantifies the level of alignment between the original and reproduced results. Finally, for Type IV, we provide the comparison of conclusions and key findings from both the original and reproduction experiments.

Type I First, we report the score human evaluation result using Best-Worst Scaling method. We report the score of original experiment and our reproduction experiment. Next, we calculated the coefficient of variation (CV) values for each model across different scenarios to assess the precision of the results. Following Belz (2022), we adjusted the CV for small sample sizes, referring to this adjusted value as CV*. Since the measurements included negative values, we shifted the measurement scale by adding 1 to ensure all values were positive, according to the recommendation of Belz (2025) regarding such shifting. The results are presented in Table 1.

Our findings are similar to the previous study, showing that NCLS is the best model in the minimum scenario, while MCLAS is the best model in the medium and maximum scenarios. However, in some results, only the maximum scenario has a low CV* score, which lower CV* score represents better result. This indicate that only the reproduction results of the maximum scenario are close to the original study.

Type II Next, we report the correlation between original and reproduction study using Pearson and Spearman correlations. The result is presented in Table 2. In the maximum scenario, both linear and monotonic relationships are nearly perfect and statistically significant. In the minimum and medium scenarios, the correlations appear strong, but they are not statistically validated, possibly due to smaller sample size.

Type III Next, we report the Krippendorff's α value from the original and reproduction annotation

results. We report almost all of the models have low values of Krippendorff's α. These shows the less agreement between original and reproduction study for each annotations.

Type IV Finally, we report whether the findings from the original experiment were verified in our reproduction study. The original study claimed three key findings, listed in Table 4. However, due to instructions from the organizers, our evaluation focused solely on the "informativeness" metric, limiting verification to claims related to this aspect. Regarding claim 1 , from the original study, both MCLAS and NCLS+MS showed improved performance as the data size increased; in our reproduction, only MCLAS was confirmed to exhibit such improvement. For claim 2 , from the original study, MCLAS outperformed both NCLS and NCLS+MS only in the maximum scenario, whereas in our reproduction, MCLAS outperformed these systems not only in maximum scenario, but also in medium scenario. Claim 3 falls outside the scope of this reproduction and could not be assessed. Overall, both the original and reproduction experiments confirm that the MCLAS model performs best among the models.

5 Discussion

From the results, we conclude that the reproduction findings align with the original study. In the minimum scenario, the best model is NCLS, while for the Medium and Maximum scenarios, the best model is MCLAS. However, the correlation scores indicate only slight agreement. We hypothesize that this may be due to annotator quality, as we recruited master's students studying NLP. If we had chosen experts in both Chinese and English language, the annotation quality might have been significantly better.

When reviewing the annotations, we noticed that some annotators occasionally scored the models inconsistently in a small occurence. For instance, in a single paragraph, two or three models output might be labeled as worst (-1) or best (1). This inconsistency arose because the annotation platform did not restrict such settings. To address this, we contacted the annotators with these issues and asked them to reannotate the data manually, providing them with the correct annotations as a reference. Surprisingly, we also found this errors in original study, where there is a participant score two models as best (1).

Additionally, upon reviewing the incorrect annotations, we suspect that the Best-Worst Scaling method may not be the most appropriate option for rating these outputs. As illustrated in Figure 2, the outputs from models 3 and 4 are both uninformative and provide incorrect information within the paragraph. This may lead to confusion for the annotators when selecting only one result to be marked as the worst. We suggest that it might be more effective to use a different approach to evaluate the models, such as rating each result on a scale from worst to best (1-5).

From these findings, we recognize the critical importance of annotator quality in achieving consistent evaluation, especially when dealing with multiple languages. We also understand that the choice of evaluation methodology significantly impacts the quality of the results.

6 Conclusion

In this study, we report our reproduction experiment from paper "Cross-Lingual Abstractive Summarization with Limited Parallel Resource". We reproduce the human evaluation with the similar setup as the original paper reported, but we only evaluate one metric instead of three by following the instructions from the organizer. By comparing the results between original and reproduction study, we found that the scores differs in several models. This highlights the importance of the choice of evaluation methodology and evaluators.

Acknowledgments

The present research was partially supported by the Key Research and Development Program of China (Grant No. 2023YFE0116400). We would like to thank Craig Thomson for the help and guidance for this experiment.

References

Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. 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 6910–6924, Online. 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 and Craig Thomson. 2024. 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, 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.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, Ling Shi, Bojian Jiang, and Deyi Xiong. 2024a. Large language model safety: A holistic survey. *CoRR*, abs/2412.17686.

Dan Shi, Chaobin You, Jiantao Huang, Taihao Li, and Deyi Xiong. 2024b. CORECODE: A common sense annotated dialogue dataset with benchmark tasks for chinese large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18952–18960. AAAI Press.

Supryadi, Leiyu Pan, and Deyi Xiong. 2024. An empirical study on the robustness of massively multilingual neural machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1086–1097, Torino, Italia. ELRA and ICCL.

Linhao Yu, Qun Liu, and Deyi Xiong. 2024. LFED: A literary fiction evaluation dataset for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10466–10475, Torino, Italia. ELRA and ICCL.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.