Rethinking Efficient Multilingual Text Summarization Meta-Evaluation

Rilyn R. Han[*1] **Jiawen Chen**[*1] **Yixin Liu**[1] **Arman Cohan**[1,2]
[1]Yale University [2]Allen Institute for AI
{rilyn.han, jiawen.chen, yixin.liu, arman.cohan}@yale.edu

Abstract

Evaluating multilingual summarization evaluation metrics, i.e., meta-evaluation, is challenging because of the difficulty of human annotation collection. Therefore, we investigate an efficient multilingual meta-evaluation framework that uses machine translation systems to transform a monolingual meta-evaluation dataset into multilingual versions. To this end, we introduce a statistical test to verify the transformed dataset quality by checking the meta-evaluation result consistency on the original dataset and back-translated dataset. With this quality verification method, we transform an existing English summarization meta-evaluation dataset, RoSE, into 30 languages, and conduct a multilingual meta-evaluation of several representative automatic evaluation metrics. In our meta-evaluation, we find that metric performance varies in different languages and neural metrics generally outperform classical text-matching-based metrics in non-English languages. Moreover, we identify a two-stage evaluation method with superior performance, which first translates multilingual texts into English and then performs evaluation. We make the transformed datasets publicly available to facilitate future research.

1 Introduction

The evaluation of automatic evaluation metrics, or *meta-evaluation*, is critical for the development and evaluation of text summarization models. However, conducting a meta-evaluation study is difficult and resource-intensive because it can require a large collection of human annotations for calculating the correlations between metric and human evaluation. This is especially true in multilingual contexts, where resources are scarce and recent advancements have predominantly centered around English. As a result, there are only a few human annotation benchmarks for summarization meta-evaluation in recent years (Huang et al., 2020; Bhandari et al., 2020; Fabbri et al., 2021; Zhang and Bansal, 2021; Gao and Wan, 2022; Liu et al., 2023), and even fewer that are multilingual (Aharoni et al., 2023; Clark et al., 2023; Hada et al., 2023).

The limited availability of multilingual meta-evaluation datasets makes it challenging to evaluate automatic metrics in multilingual settings. Therefore, we aim to explore an efficient multilingual meta-evaluation strategy based on monolingual data. Specifically, inspired Braun et al. (2022), we revisit and explore the possibility of transforming a monolingual (English) meta-evaluation dataset into a multilingual dataset through neural machine translation (NMT), which can avoid the difficulty of collecting human annotations in different languages, especially the low-resource languages. To this end, our study includes two main parts:

We first investigate the soundness and feasibility of our NMT-based dataset transformation approach. Specifically, we identify the key question to be **whether the translated dataset is of good quality and supports meaningful multilingual summarization meta-evaluation**. To verify this question, we propose a dataset quality measurement strategy based on neural back-translation, or round-trip translation (Mallinson et al., 2017). Specifically, we hypothesize that if a translated dataset is of good quality, then its back-translated dataset in English should yield ***consistent*** meta-evaluation results as the original English dataset for the same automatic metric. To measure this consistency, we adopt a statistical test, Two One-Sided Test (TOST) (Schuirmann, 1987), which allows us to assess this consistency for our use case. While Braun et al. (2022) also applies TOST for checking dataset quality, our test is fundamentally different since they aim to check whether metrics perform similarly on the translated dataset, while we focus on the back-translated dataset, which avoids the metric performance discrepancy in different

[*] Equal contribution

Findings of the Association for Computational Linguistics ACL 2024, pages 15739–15746
August 11-16, 2024 ©2024 Association for Computational Linguistics

languages. Furthermore, to better understand the effect of translation quality in our task, we compare two NMT systems for dataset transformation, including a domain-specific NMT model, M2M-100 (Fan et al., 2021), and a generic large language model (LLM), GPT-3.5 (Ouyang et al., 2022). Our results indicate that translation quality is crucial in our context, and the data quality testing method we propose is sensitive to the performance of the translation systems employed during our dataset construction.

After verifying the soundness of our proposed dataset transformation approach, we conduct a multilingual summarization meta-evaluation on the dataset transformed into 30 languages from a recently introduced English summarization meta-evaluation dataset, RoSE (Liu et al., 2023). This results in a much larger dataset compared with previous work (Braun et al., 2022) in terms of both the number of languages (30 v.s. 7) and data examples in the datasets. To ensure meaningful evaluation results, we use the TOST results we obtained as a filtering criterion – we accept only the evaluation results that can pass the consistency test we proposed. Our meta-evaluation highlights the metric performance discrepancy in different languages and reveals their limitations. Furthermore, we identify a two-stage automatic evaluation approach, which first translates multilingual texts into English and then applies automatic metrics to the translated text, achieving better average performance than applying the metrics to multilingual texts directly.

Our contributions are three-fold:

(1) We conduct a thorough analysis of an efficient meta-evaluation method using NMT-based data transformation and design a statistical test procedure to verify the soundness of this approach.

(2) We conduct a multilingual meta-evaluation in 30 languages, shedding light on the limitations of current summarization evaluation metrics.

(3) We release the transformed datasets in all languages to facilitate future research.[1]

2 Dataset Creation

To circumvent the difficulty of collecting multilingual human annotations, we apply a data creation strategy transforming existing monolingual datasets into a multilingual one using NMT systems. Specifically, we choose RoSE (Liu et al., 2023), a recently introduced summarization meta-evaluation dataset for our study. RoSE contains human annotations for reference-based summary salience evaluation based on a fine-grained protocol, Atomic Content Units (ACUs), which aims to reduce the subjectivity of reference-based summarization human evaluation. We use its subset from the CNN/DailyMail (Nallapati et al., 2016) test set in our experiments since it contains longer summaries with higher complexity. It contains 500 distinct documents, each accompanied by 12 human-annotated scores of system-generated summaries.

ISO	Lang.	ISO	Lang.	ISO	Lang.
AR	Arabic	GA	Irish	PL	Polish
BG	Bulgarian	HE	Hebrew	PT	Portuguese
BN	Bengali	HI	Hindi	RU	Russian
CS	Czech	HU	Hungarian	SV	Swedish
DE	German	ID	Indonesian	TA	Tamil
EL	Greek	IT	Italian	TR	Turkish
ES	Spanish	JA	Japanese	UK	Ukrainian
FA	Persian	KO	Korean	VI	Vietnamese
FI	Finnish	LT	Lithuanian	YI	Yiddish
FR	French	NL	Dutch	ZH	Chinese

Table 1: 30 languages we used in the meta-evaluation. For each language, we display the ISO code.

We use two NMT systems to transform RoSE into 30 languages: (1) M2M-100, a multilingual encoder-decoder model primarily intended for translation tasks, which is chosen because it supports a wide range of languages; (2) GPT-3.5 (Ouyang et al., 2022),[2] an LLM that has shown strong performance on machine translation especially when translating from and to a *pivot* language, e.g., English (Jiao et al., 2023). The 30 languages we selected are listed in Table 1. Representing 16 diverse language families, these languages ensure a thorough linguistic range.

To evaluate the translation quality, we use round-trip translation to translate the translated dataset back to English. We then use a few standard automatic metrics to compare the similarity between the back-translated dataset and the original English dataset. We generally observe a high variance in performance, with mostly resource-rich languages like German, Italian, and French achieving a high score (e.g., Rouge-2 of 70+%), while some other languages achieving lower performance (e.g., Bengali, Japanese and Tamil with Rouge-2 of 30% or less). The full results are in Appendix A.

[1]The datasets are available at https://github.com/yale-nlp/MRoSE.

[2]gpt-3.5-turbo is used, described in https://platform.openai.com/docs/models/gpt-3-5.

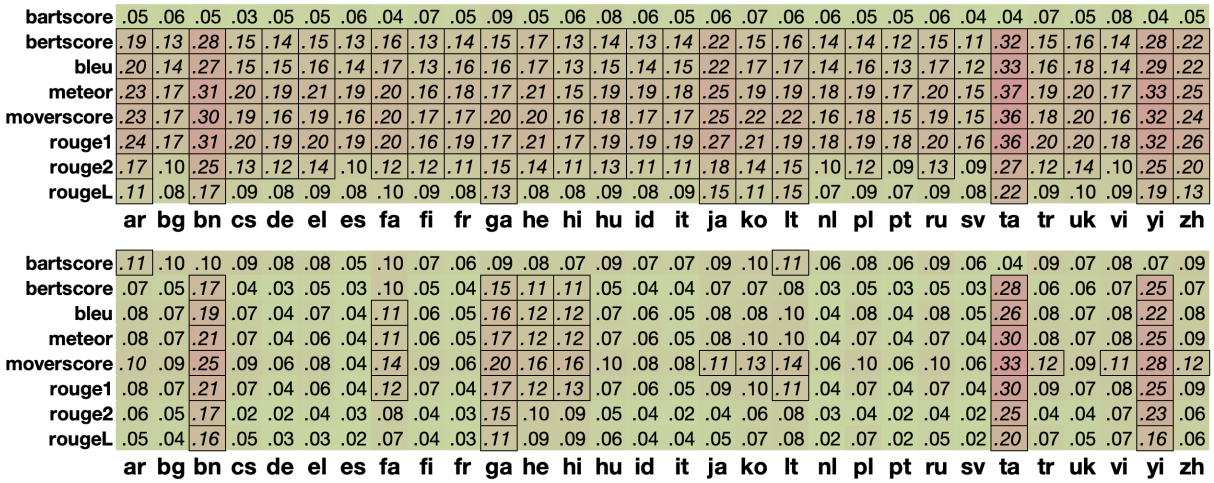

Figure 1: TOST results on the multilingual datasets translated by M2M-100 (top) and GPT-3.5 (bottom). The correlation equivalence margins between the original dataset and the back-translated English dataset are reported. The data points that fail to pass the pre-defined margin (0.1) are highlighted with bounding boxes.

3 Dataset Analysis

We design and conduct statistical analyses to verify whether our dataset obtained through machine translation can support reliable evaluation of automatic evaluation metrics.

Our dataset quality verification process relies on a key assumption – the meta-evaluation performed on the original dataset and the back-translated dataset should yield ***consistent*** results if the dataset transformation is reliable. With this assumption, we leverage the Two One-Sided Test (TOST) (Schuirmann, 1987) to check this consistency property, which is frequently utilized to test whether the difference between two groups is negligible. Specifically, a small difference in meta-evaluation results between the original dataset and its back-translated counterpart can indicate the maintained quality in our translated dataset. Here, the meta-evaluation results are the correlations between the metric and the human evaluation results (Bhandari et al., 2020; Deutsch et al., 2021).[3] Therefore, we apply TOST to test the equivalence of the human-metric correlations on the original dataset (ρ_{orig}) and the back-translated dataset (ρ_{back}). To execute TOST, we formulated both null and alternative hypotheses:

$$H_0 : |\rho_{\text{orig}} - \rho_{\text{back}}| > \Delta_e, \quad H_1 : |\rho_{\text{orig}} - \rho_{\text{back}}| < \Delta_e \quad (1)$$

The null hypothesis, represented by H_0, assumes that the absolute difference between two correlations cannot be ignored when considering a pre-defined equivalence margin Δ_e. Conversely, the alternative hypothesis (H_1) suggests that the difference is small enough to infer that the two correlations are nearly identical. By conducting TOST, our goal is to identify the equivalence margin (Δ_e) at which we can reject the null hypothesis and conclude that two correlations are sufficiently close. For each automatic metric we are going to evaluate, we execute TOST with bootstrapping (Tibshirani and Efron, 1993) on the datasets translated into 30 languages with Kendall rank correlation as the correlation coefficient, and search for the margin Δ_e by identifying the point at which the null hypothesis can be rejected ($p < 0.05$).

We report the TOST results on the datasets translated by M2M-100 and GPT-3.5 in Figure 1. For M2M-100, we find that the majority of the meta-evaluation results cannot be regarded as equivalent under an equivalence margin (Δ_e) of 0.1. On the other hand, for GPT-3.5, the majority of the results are considered equivalent with the same margin.

The test provides a systematic way to assess the translated dataset quality and highlights the importance of using a strong NMT system for the dataset transformation. Therefore, we propose a data-filtering method for the actual multilingual meta-evaluation – with a pre-defined margin (Δ_e), the meta-evaluation data point (i.e., one metric evaluated on one language) that cannot pass TOST will be discarded. As a case study, we set this margin to 0.1 in our study, noting that it can be adjusted based on the level of error tolerance.

[3] We report the summary-level correlations since it is more challenging than system-level correlations (Liu et al., 2023).

Figure 2: Automatic metric performance (Kendall's correlation with human evaluation) on subsets in different languages translated by GPT-3.5. We report metric performance on the multilingual text and the back-translated English text, the latter denoted by the suffix `_bt`. The invalid results that failed to pass TOST are grayed out.

4 Evaluating Multiligual Automatic Summarization Evaluation

We now conduct a meta-evaluation of different multilingual metrics on the dataset we obtained using GPT-3.5 as the NMT system.

4.1 Metrics

We evaluate a few widely used automatic metrics for evaluating textual similarity, including text-matching based metrics: (1) ROUGE-1,2,L (Lin, 2004), (2) BLEU (Papineni et al., 2002), (3) METEOR (Banerjee and Lavie, 2005); and neural metrics: (1) BERTScore (Zhang et al., 2019b), (2) BARTScore (Yuan et al., 2021), (3) MoverScore (Zhao et al., 2019). For each neural metric, we use their multilingual version powered by base multilingual language models. Specifically, we use the `bert-base-multilingual-cased` version for BERTScore and MoverScore, and `mbart-large-50` for BARTScore which is based on MBART (Tang et al., 2021).[4] We note that since RoSE contains system summary *recall* scores against the reference summary, we use the recall score of the metrics when available.

We also investigate a two-stage automatic evaluation strategy – using an NMT model to translate the text into English first, then applying the metrics to the translated English text. Therefore, for each metric we also evaluate their performance on English text translated by GPT-3.5.

4.2 Meta Evaluation

We evaluate the metric performance on the translated datasets by calculating the correlation between the metric scores and human-annotated scores. As mentioned above, we use the TOST results we obtained in §3 to filter out the invalid evaluation results. The results on the datasets translated by GPT-3.5 are in Figure 2, and the results with M2M-100 translations can be found in Appendix B. We note several key findings:

(1) **Metric performance varies across different languages**, and in general they perform worse on non-alphabetical languages. For example, most metrics achieve worse performance in Chinese (zh) than in German (de), indicating their limitations in multilingual evaluation settings.

(2) **Different metrics show large performance gaps on certain languages**. Specifically, text-matching based metrics such as ROUGE and BLEU are less robust than neural metrics such as BERTScore. We believe this is because neural metrics based on multilingual language models can better capture the semantic similarity than ROUGE.

(3) **NMT-based two-stage evaluation methods achieve strong performance**. Specifically, all metrics we evaluated tend to perform better when evaluating the translated English summaries than evaluating the original summaries in different languages, with a few exceptions such as BERTScore in Chinese (zh) or ROUGE-2 in Vietnamese (vi).

5 Conclusion

In this work, we investigate an efficient multilingual summarization meta-evaluation approach by translating a monolingual dataset into a multilingual version. We develop a statistical testing procedure to verify the transformed dataset quality to ensure the validity of the meta-evaluation results. With this procedure, we compare the dataset quality translated by two NMT systems, and we found that the translation quality is critical for multilin-

[4]The model checkpoints are from Hugging Face Models: https://huggingface.co/models.

gual meta-evaluation. We then present a case study of transforming a monolingual meta-evaluation dataset into 30 languages with high-quality translations, and conduct a multilingual meta-evaluation of representative automatic evaluation metrics on the transformed dataset.

6 Limitations

Our work utilizes the NMT systems to transform a monolingual meta-evaluation dataset into a multilingual version. While we have developed statistical tests to evaluate the quality of the translated dataset, we recognize that it is not a replacement for human evaluation of the dataset quality. However, we found it difficult to conduct such human evaluation in multilingual settings across many languages, which we leave for more comprehensive future work. Similarly, in our meta-evaluation study, we only report the quantitative metric performance by computing their correlations with the human evaluation results. We believe that a qualitative study with human evaluation would provide a more in-depth understanding of the metric performance and behavior, which is an important next step to reliability extend the summarization meta-evaluation research into multilingual settings.

We acknowledge that there are other metrics we did not include in our meta-evaluation, which could have made this meta-evaluation study more complete. Besides, we only investigate one quality dimension, the summary salience. It is possible that our data construction process can have a different impact on the other quality dimensions such as summary coherence or factuality.

We note that the (back-)translation technique we employed can lead to several concerns. First, our translated multilingual dataset may have different distributions/characteristics compared to datasets that are originally multilingual mostly because of issues in translationese (Wang et al., 2023). Moreover, since we did not employ human evaluations to check the translated dataset, it is possible that some of the translation errors (e.g., failing to translate from English to another language) are not captured in our data quality verification method.

Acknowledgements

We thank the anonymous reviewers for their invaluable comments and suggestions. We are grateful to the OpenAI's Researcher Access Program for API credit support.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.

Spencer Braun, Oleg Vasilyev, Neslihan Iskender, and John Bohannon. 2022. Does summary evaluation survive translation to other languages? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2425–2435, Seattle, United States. Association for Computational Linguistics.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur P Parikh. 2023. Seahorse: A multilingual, multifaceted dataset for summarization evaluation. *arXiv preprint arXiv:2305.13194*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav

Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *arXiv preprint arXiv:2309.07462*.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough? In *NTCIR*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

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.

Donald J Schuirmann. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15:657–680.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436.

Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay Cohen, Manish Shrivastava, and Barry Haddow. 2023. PMIndiaSum: Multilingual and cross-lingual headline summarization for languages in India. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11606–11628, Singapore. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023. Understanding translationese in cross-lingual summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3837–3849, Singapore. Association for Computational Linguistics.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. Ted: A pretrained unsupervised summarization model with theme modeling and denoising. *arXiv preprint arXiv:2001.00725*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019a. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

A Translation Quality

In §2, we apply two NMT systems to transform the English meta-evaluation dataset into multilingual datasets. To have an initial check of the translation quality, in Table 2 and Table 3 we report the metric scores between the original English reference summaries and the back-translated English reference summaries.

B Additional Meta-Evaluation Results

In §4.2, we report the meta-evaluation results on the datasets translated by GPT-3.5. As a reference, in Figure 3 we report the meta-evaluation results on datasets translated by M2M-100. We note that since most of the subsets translated by M2M-100 cannot pass our equivalence test, TOST, the meta-evaluation results on these subsets are unreliable.

Language (Code)	BLEU	ROUGE-1	ROUGE-2
Arabic (AR)	30.37	63.11	36.81
Bulgarian (BG)	46.91	69.90	46.35
Bengali (BN)	15.88	53.29	28.87
Czech (CS)	54.60	72.45	50.34
German (DE)	49.90	76.83	58.07
Greek (EL)	58.58	70.37	48.29
Spanish (ES)	65.34	76.01	55.52
Persian (FA)	51.89	64.68	38.48
Finnish (FI)	44.24	70.44	46.40
French (FR)	59.51	75.06	54.11
Irish (GA)	41.76	66.21	44.01
Hebrew (HE)	33.81	65.16	40.30
Hindi (HI)	56.90	70.28	44.51
Hungarian (HU)	43.28	69.09	43.91
Indonesian (ID)	52.83	73.84	50.21
Italian (IT)	58.72	75.11	53.99
Japanese (JA)	18.36	60.61	33.44
Korean (KO)	46.79	65.08	39.31
Lithuanian (LT)	61.76	65.00	41.77
Dutch (NL)	46.13	78.34	61.03
Polish (PL)	46.25	70.17	46.37
Portuguese (PT)	53.19	77.11	56.43
Russian (RU)	50.29	66.89	42.22
Swedish (SV)	60.33	79.63	63.18
Tamil (TA)	66.03	43.51	35.26
Turkish (TR)	37.93	67.61	41.51
Ukrainian (UK)	48.34	65.55	40.78
Vietnamese (VI)	41.23	72.11	49.48
Yiddish (YI)	19.58	57.82	45.36
Chinese (ZH)	26.32	62.76	35.35

Table 2: BLEU and ROUGE scores between the original English reference summaries and the back-translated English reference summaries using M2M-100 across different languages.

C Additional Related Work

Summarization Evaluation Neural text summarization has seen great success in recent years (Liu and Lapata, 2019; Zhang et al., 2019a; Yang et al., 2020). Evaluation metrics like ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005) traditionally assess n-gram overlap, while neural metrics such as BERTScore (Zhang et al., 2019b), MoverScore (Zhao et al., 2019), and BARTScore (Yuan et al., 2021) utilize neural language models for similarity measurement. To verify the effectiveness of such summarization metrics, many related studies conduct meta-evaluation, which involves calculating the correlation between the automatic metric and human-annotated scores (Bhandari et al., 2020;

Figure 3: Automatic metric performance on subsets in different languages translated by M2M-100. Kendall's correlation between metric evaluation and human evaluation results is reported. For each metric, we report its performance on the multilingual text and the back-translated English text, the latter denoted by the suffix `_bt`. The invalid results that failed to pass TOST are grayed out.

Language (Code)	BLEU	ROUGE-1	ROUGE-2
Arabic (AR)	67.37	69.82	43.71
Bulgarian (BG)	58.88	74.10	50.05
Bengali (BN)	23.47	56.72	28.07
Czech (CS)	56.49	74.72	49.97
German (DE)	87.30	78.21	56.28
Greek (EL)	44.62	74.58	51.22
Spanish (ES)	72.42	80.17	59.60
Persian (FA)	37.07	65.31	37.42
Finnish (FI)	65.92	74.46	50.59
French (FR)	56.66	77.33	55.42
Irish (GA)	27.61	61.74	34.93
Hebrew (HE)	24.59	65.32	39.02
Hindi (HI)	43.76	67.36	39.88
Hungarian (HU)	59.16	72.18	46.64
Indonesian (ID)	68.57	74.94	50.80
Italian (IT)	74.91	79.42	58.54
Japanese (JA)	47.45	68.16	39.89
Korean (KO)	20.95	66.09	37.19
Lithuanian (LT)	20.27	70.33	44.66
Dutch (NL)	80.15	79.93	59.24
Polish (PL)	57.81	75.25	51.09
Portuguese (PT)	74.47	79.43	58.43
Russian (RU)	50.44	71.86	46.40
Swedish (SV)	71.79	79.49	59.42
Tamil (TA)	5.88	36.88	11.57
Turkish (TR)	42.06	72.58	46.48
Ukrainian (UK)	66.92	70.86	45.12
Vietnamese (VI)	70.90	70.29	44.59
Yiddish (YI)	17.09	51.20	27.64
Chinese (ZH)	25.10	68.19	40.23

Table 3: BLEU and ROUGE scores between the original English reference summaries and the back-translated English reference summaries using GPT-3.5 across different languages.

Huang et al., 2021; Fabbri et al., 2021; Pagnoni et al., 2021; Zhang and Bansal, 2021; Gao and Wan, 2022; Liu et al., 2023). Most of these studies are conducted on English corpora only, which is likely because of the limitations of datasets, summarization systems and metrics, and the difficulty of conducting multilingual human evaluation.

Multilingual Summarization Datasets Despite the rapid progress of text summarization research, the availability of multilingual text summarization datasets remains limited. MLSUM (Scialom et al., 2020) is a pioneering multilingual summarization dataset with 1.5 million article-summary pairs in five languages (French, German, Spanish, Russian, and Turkish). Another dataset, XLSUM (Hasan et al., 2021), contains news articles from the BBC news website and the associate summaries in 45 languages. WikiLingua (Ladhak et al., 2020) is a large-scale dataset focusing on cross-lingual abstractive summarization, with 141,000 English articles and parallel articles for 17 languages. There are even fewer multilingual summarization meta-evaluation datasets. Among them, mFACE (Aharoni et al., 2023) evaluates summarization faithfulness on XLSum across 45 languages and releases the human annotations. SEAHORSE (Clark et al., 2023), another meta-evaluation dataset, includes 96,000 summaries across 6 languages and 9 summarization systems with multi-dimensional human annotations of summary quality. Hada et al. (2023) conducted a meta-evaluation of LLM-based evaluators under multilingual settings in 9 languages. They found that while LLMs can achieve high agreements with human annotators on *reference-free* summarization evaluation, the LLMs' performance varies across different languages. There are also more recent cross-lingual summarization datasets, including CrossSum (Bhattacharjee et al., 2023) and PMIndiaSum (Urlana et al., 2023).