Rewriting the Narrative: Evaluating Machine Translation in African Languages with Corrected FLORES Data

A Comparative Study on the Impact of Dataset Corrections on Translation Accuracy and Model Fairness

Itumeleng Moshokoa, Bukhosi Eugene Mpande, Ennis Maphasha
Group Number 24

Introduction

The FLORES datasets are widely used to benchmark machine translation (MT) in low-resource languages. However, inconsistencies and translation errors in the original versions have raised concerns about the validity of past evaluations. Targeted corrections for four African languages—Hausa, Northern Sotho, Xitsonga, and isiZulu—aim to improve evaluate accuracy and fairness, of the improved and corrected models using the previous versions as a benchmark [1].

Research Questions & Objectives

Central Research Question

How do the corrections to the FLORES dataset influence the accuracy, fairness, and reliability of machine translation model evaluations for low-resource African languages?

Sub-Questions

- **Model Performance and Retraining:** To what extent do the corrections affect the performance rankings of existing machine translation models, and is it more effective to retrain these models from scratch or fine-tune them using the corrected data?
- **Translation Quality and Error Impact:** What changes in translation accuracy, quality, and efficiency are observed with the corrected dataset, and which types of errors in the original version had the most significant impact on model evaluation outcomes?
- **Future-Proofing Dataset Evaluation:** What validation methods and best practices can be established to ensure that future low-resource language datasets support fair, accurate, and robust benchmarking in NLP?

Methodology

This project adopts a comparative evaluation approach to assess how corrections in the FLORES dataset affect machine translation (MT) performance for four African languages: Hausa, Northern Sotho, Xitsonga, and isiZulu. We will benchmark pre-trained multilingual models (e.g., mBART, NLLB) on both the original and corrected FLORES dev/devtest splits to isolate the effect of dataset quality on translation accuracy and fairness.

Performance will be measured using BLEU, CHrF, COMET, and BERTScore. Changes in model rankings and error types will be tracked to quantify the impact of corrections. To gain deeper insights, we will apply explainability methods such as attention visualization and error clustering.

Robustness Testing

We will also evaluate model performance across dialects and domains (e.g., legal, conversational) using FLORES+ metadata to assess generalization and real-world applicability.

This combined quantitative and qualitative strategy ensures a well-rounded assessment of how dataset quality influences MT outcomes for low-resource languages.

Datasets Used

- **FLORES-101**
  - Size: 3,001 professionally translated sentences per language across 101 languages.
  - Splits: dev (997 sentences) and devtest (1,012 sentences).
  - Attributes: `id`, `sentence`, `URL`, `domain`, `topic`, `has_image`, `has_hyperlink`.
  - Source: Available via Hugging Face. [2]
- **FLORES-200**
  - Size: 3,001 sentences translated into 200 languages, drawn from 842 distinct web articles.
  - Attributes: `id`, `sentence`, `URL`, `domain`, `topic`, `has_image`, `has_hyperlink`.
  - Source: Available via GitHub. [3] [4] [5]
- **FLORES+**
  - Size: 1,000 sentences per language per split, based on FLORES-200.
  - Attributes: `id`, `text`, `iso_639_3`, `iso_15924`, `glottocode`, `URL`, `domain`, `topic`, `has_image`, `has_hyperlink`, `last_updated`.
  - Source: Available via Hugging Face. [6]
- **FLORES Fix 4 Africa**
  - Size: Approximately 2,009 manually corrected sentences for Hausa, isiZulu, Northern Sotho, and Xitsonga.
  - Attributes: `id`, `sentence`, `URL`, `domain`, `topic`, `has_image`, `has_hyperlink`.
  - Source: Available via GitHub. [7] [1]

All four datasets are publicly accessible and well-documented, requiring no additional data collection.

Evaluation Strategy

To evaluate the effectiveness of machine translation models, we will compare performance on both the original and corrected FLORES datasets for four African languages. This original dataset serves as the baseline, while the corrected version provides the experimental benchmark.

The following evaluation metrics will be used:

- **BLEU (Bilingual Evaluation Understudy)** – Measures n-gram overlap between model output and reference translations.
  Success Indicator: Higher BLEU scores indicate more accurate word-level translation.
- **chrF (Character F-score)** – Captures character-level precision and recall.
  Success Indicator: Higher chrF scores reflect better performance in morphologically rich African languages.
- **COMET** – A learned evaluation metric aligned with human judgment.
  Success Indicator: Higher COMET scores suggest closer alignment with human-rated translation quality.
- **BERTScore** – Evaluates semantic similarity using contextual embeddings.
  Success Indicator: Higher scores indicate better preservation of meaning and context.
- **Error Type Frequency Distribution** – Tracks common translation issues such as omissions or hallucinations.
  Success Indicator: Fewer critical errors and improved fluency.
- **Model Ranking Stability** – Compares ranking consistency across datasets.
  Success Indicator: Meaningful rank shifts reveal the impact of dataset corrections; high consistency indicates robust model behavior.

Benchmarks: Evaluation will use the original FLORES-101/FLORES-200 datasets as baselines and the corrected FLORES Fix 4 Africa as the benchmark for improvement. FLORES+ will support robustness checks across domains and dialects.

Expected Outputs and Contributions

This study aims to uncover how dataset quality—specifically through dataset corrections—impacts translation model performance and fairness. The following are the expected areas of analysis and insight:

- **Significant Changes in Model Rankings:** Evaluating models on corrected datasets is

expected to reveal significant shifts in model rankings compared to uncorrected data. This highlights the critical role of high-quality evaluation data in producing reliable and fair performance assessments.
Deliverable: Performance deltas (e.g., BLEU, COMET) for each language pair showing how rankings change post-correction.

- **Error Analysis Report:** A detailed analysis will be conducted to identify common types of dataset errors—such as mistranslations, inconsistent phrasing, or domain mismatches—that disproportionately affect evaluation scores.
Deliverable: Categorized list of error types, frequency distribution, and their measurable impact on translation quality metrics.
- **Model Sensitivity to Dataset Quality:** Comparative analysis will be performed to understand which types of translation models (e.g., transformer-based, multilingual vs. bilingual, low-resource vs. high-resource fine-tuning) are most affected by dataset imperfections.
Deliverable: Insights into architecture-level sensitivity to dataset quality, supported by quantitative performance variance across corrected vs. original datasets.

References

[1] I. Abdulmumin, S. Mkhwanazi, M. S. Mbooi, S. H. Muhammad, I. S. Ahmad, N. Putini, M. Mathebula, M. Shingange, T. Gwadabe, and V. Marivate, "Correcting flores evaluation dataset for four african languages," 2024. [Online]. Available: https://arxiv.org/abs/2409.00626

[2] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The Flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.30/

[3] J. C. O. M. E. K. H. K. H. E. K. J. L. D. L. J. M. A. S. S. W. G. W. A. Y. B. A. L. B. G. M. G. P. H. J. H. S. J. K. R. S. D. R. S. S. C. T. P. A. N. F. A. S. B. S. E. A. F. C. G. V. G. F. G. P. K. A. M. C. R. S. S. H. S. J. W. NLLB Team, Marta R. Costa-jussà, "No language left behind: Scaling human-centered machine translation," 2022.

[4] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," 2021.

[5] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english," 2019.

[6] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "Scaling neural machine translation to 200 languages," *Nature*, vol. 630, no. 8018, pp. 841–846, 2024. [Online]. Available: https://doi.org/10.1038/s41586-024-07335-x

[7] I. Abdulmumin, S. Mkhwanazi, M. Mbooi, S. H. Muhammad, I. S. Ahmad, N. Putini, M. Mathebula, M. Shingange, T. Gwadabe, and V. Marivate, "Correcting FLORES evaluation dataset for four African languages," in *Proceedings of the Ninth Conference on Machine Translation*, B. Haddow, T. Kocmi, P. Koehn, and C. Monz, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 570–578. [Online]. Available: https://aclanthology.org/2024.wmt-1.44/