|  |
| --- |
| **Impact of Reference Corrections on Machine Translation Model Rankings in Low-Resource African Languages** |
|  |
| **Alec Watson (u22491351), Aren Repko (u04515791), Daniel Geerdink (u22556860)** |
|  |
|  |
|  |
|  |

Abstract

Accurate evaluation is critical for advancing machine translation (MT) in low-resource languages, yet flawed reference translations can distort model performance metrics and rankings. This study investigates the impact of corrected references—using the recently improved FLORES Fix for Africa dataset—on the evaluation of MT systems for four African languages: Hausa, isiZulu, Xitsonga, and Northern Sotho. We assess model performance across standard metrics and analyse shifts in rankings, score sensitivity, and domain-specific effects. Our results show minimal changes in model rankings despite slight score improvements with corrected references, likely due to their inclusion in recent model training. NLLB-200 outperforms larger models, highlighting the value of language-specific optimisation over scale. Domain and language disparities persist, underscoring the need for better support and evaluation practices for low-resource African languages.

Introduction

Machine Translation (MT) for African languages remains significantly under-resourced and under-evaluated, especially when compared to high-resource languages. While datasets such as FLORES-200 have enabled benchmarking for African languages like Hausa, isiZulu, Northern Sotho, and Xitsonga, recent research has revealed that these test sets contain notable errors (Abdulmumin et al., 2024). These errors undermine the fairness and validity of model evaluations, especially in low-resource languages like isiZulu, Hausa, Xitsonga, and Northern Sotho. There is a lack of systematic analysis on how corrected references impact the relative performance of translation models. How do manual corrections to reference datasets affect the reliability of MT evaluations, and what role do model architecture and textual domains play in these outcomes? We hope answers to this question (and others like it) will help the quest for linguistic equity allowing speakers of all languages access to information and the digital ecosystem.

We hypothesize that the use of corrected reference translations will result in measurable changes in automatic evaluation scores (BLEU, COMET, BERTScore will drop using the corrected data set), but will not significantly alter the relative rankings of the evaluated machine translation models. Furthermore, we expect that score sensitivity to these corrections will vary across textual domains, with entertainment and informal domains being more affected than formal ones (e.g., government, religion).

Literature Survey

Recent advances in machine translation (MT) particularly with models like NLLB (Team et al., 2022), and OPUS-MT (Tiedemann & Thottingal, 2020) and MADLAD-400 (Kudugunta, S. et al. 2023) have improved multilingual performance but not the resource scarcity faced by African Languages. The FLORES-200 dataset (No Language Left Behind: Scaling Human-Centered Machine Translation) has been widely adopted for evaluating models in the NLP space and specifically MT and NLP in low resource languages such as African languages. However, Abdulmumin et al. (2024) revealed that this dataset and likely many like it (due to how it was obtained) contain many errors which can and will affect the performance of models due to the dataset’s position as the standard for testing and training. In Abdulmumin et al. (2024) the dataset was corrected for Hausa, isiZulu, Xitsonga, and Northern Sotho through manual intervention of fluent speakers of the languages and the difference between the original and corrected dataset was measured in BLEU and COMET score. Even though the study was thorough in showing the number of tokens/sentences which were corrected, it also left some things unanswered. Namely did these corrections shift ranking in the perceived quality of leading MT models and did these corrections affect certain textual domains more than others?

Approach and Methodology

Data Sources and Preprocessing

Our study focuses on four African languages included in the FLORES benchmark: Hausa (hau), isiZulu (zul), Northern Sotho (nso), and Xitsonga (tso). We utilise two parallel datasets:

**FLORES-200 (original references)** from the flores-fix-4-africa repository (Abdulmumin et al., 2024)

**Corrected FLORES references** from the flores-fix-4-africa repository (Abdulmumin et al., 2024), which contain human-validated improvements to grammar, idiomatic expressions, and fluency.

Both data sets were tokenised using standard tools compatible with each model and no further text processing was applied.

* 1. Model selection and Setup

We evaluate outputs from three publicly available MT models, selected to represent architectural and training diversity:

**NLLB-200**: A massively multilingual transformer trained to support two hundred languages with a focus on low-resource inclusion.

**MADLAD-400**: Meta AI’s newest multilingual model supporting over four hundred languages, trained on Common Crawl and curated low-resource corpora, emphasising data quality and coverage.

**OPUS-MT**: A set of MarianMT models trained on domain-specific data from OPUS, commonly used for practical applications.

These models were selected for their **architectural diversity**, **language coverage**, and relevance to low-resource MT research.

.

Evaluation Procedure

We compare model outputs against both original and corrected references using three widely accepted automatic evaluation metrics:

BLEU: n-gram precision-based score.

COMET: learned metric leveraging contextual embeddings.

BERTScore: semantic similarity score using pretrained language models.

Our analysis includes:

Average score changes per model and metric.

Ranking shifts, measured via Spearman’s rankcorrelation to assess consistency.

Domain-level analysis, where test data is segmented by topic (e.g., news, religion, government) to examine how corrections affect specific content areas. It is important to note that due to the limits of the size of the data set not all domains are equally sized.

3.4 Responsible NLP and Ethical Considerations

All datasets are publicly available under open research licenses. The corrected reference translations were developed through native-speaker annotation, enhancing cultural and linguistic fidelity.

We recognise limitations in data diversity and domain coverage. Domain segmentation is approximate and may affect fine-grained conclusions. Furthermore, automatic metrics, though widely used, may not fully reflect translation adequacy in morphologically rich or underrepresented languages.

This work seeks to promote **fair and robust evaluation** practices in MT for African languages and contribute to broader efforts in **digital inclusion** and **linguistic equity**. By analysing model behaviour under corrected conditions, we aim to support the development of more trustworthy and socially responsible NLP systems.

Experiments and Results

*See all figures appended at the end of the report*

Our results in Figure 1 indicate no notable shift in model ranking from the original to the improved data set. All models appeared to perform at the same level relative to each other.

The difference in the scores visible in Figure 1 across all metrics are far smaller than those observed in Abdulmumin et al. (2024) (we observe a difference in score of about 0.5% to 0.2% as opposed to a 16% change). We believe this is due to the incorporation of the corrections into the data set since the publication of Abdulmumin et al. (2024) and models subsequent training on the newly corrected data set.

Our overall results show that the NLLB-200-distilled-600M (No language left behind model) performed the best across all languages on average. MADLAD400-3b-mt performed worst overall. This is interesting as it was the highest parameter model and featured the most languages (roughly 400 to NLLB’s 200) indicating that language flexibility plays a larger role in accuracy than parameter count.

When looking at the performance of the models with regard to the languages in Figure 1 we see a trend where the models consistently perform best when translating English to Hausa. (NLLB and MADLAD). This is somewhat expected as Hausa is the highest resource language of the evaluated languages, which reinforces the idea that under-resourced languages have worse model performance.

In terms of textual domains, in Figure 2 we observe that on average models perform worse on the entertainment domain. This tracks with the general lack of the recreational content available in African languages. Topics such as health politics and travel perform well as these topics usually have vocabularies that translate well across languages.

Figure 4 indicates the Spearman Rank correlation for our rankings showing they are highly correlated with a p-value showing high statistical significance (p < 10⁻⁷)

Reflections and Discussion

We faced a series of challenges most of which were related to the inequality of NLP for African Languages. Many models did not support all four of the languages we aimed to evaluate, and of those that did, there were frequent inconsistencies in their usage of language codes and documentation making it very hard to find the models that support the languages in question. This is a symptom of the broader problems with NLP equity and indicates there is no mainstream effort to translate African languages.

Despite efforts to ensure consistency, several methodological limitations affect the interpretability of our results. First, while we use standardised automatic metrics (BLEU, COMET, BERTScore), these metrics may not capture nuanced improvements in translation quality—especially for morphologically rich or agglutinative languages like isiZulu. Human evaluation would have provided a more holistic view of translation adequacy, particularly in low-resource contexts where semantic fidelity is harder to measure automatically.

Second, the models we selected vary not only in architecture (transformer size, multilingual training strategies) but also in training data and domain exposure. For example, NLLB-200 benefits from training specifically tailored to low-resource languages, whereas MADLAD-400 is optimised for broader multilingual coverage. The relative underperformance of MADLAD despite its larger parameter count suggests that sheer scale does not compensate for lack of targeted linguistic representation.

Third, dataset corrections may already be partially incorporated into some models due to overlapping training data. This introduces a temporal bias, making it difficult to isolate the effect of corrected references. Since FLORES Fix was released prior to some model updates, our comparisons between “original” and “corrected” data may understate the true impact corrections had at the time of release.

Even though the experiments did not require model training the amount of compute needed, to perform translations and evaluations on those translations, was still immense. This may pose a problem in replicating results and in future testing models as models trend upwards in size of parameters and datasets.

From our analysis a study like this (or this one for that matter) is time sensitive to the proposed changes in datasets, as these are likely to be incorporated and used in training. This of course will make determining the effect of the corrections difficult. This is most likely a contributing factor as to why our differences in scores from the corrected and original dataset are smaller than those observed in Abdulmumin et al. (2024).

Conclusion

This study set out to evaluate the impact of corrected reference translations on machine translation (MT) performance for four African languages—Hausa, isiZulu, Northern Sotho, and Xitsonga—using three multilingual MT models: NLLB-200, MADLAD-400, and OPUS-MT. Despite expectations of notable performance differences between original and corrected references, our results reveal minimal changes in model rankings and evaluation scores across metrics, in stark contrast to prior findings. We attribute this to the likely incorporation of corrected data into model training pipelines since the release of the FLORES corrections.

Among the models, NLLB-200 consistently outperformed others, especially in higher-resource (among the languages tested Hausa has an estimated ~80 million speakers to Zulu’s ~13 million) language pairs such as English Hausa, reinforcing the persistent gap in MT quality for more under-resourced languages. Domain-specific analysis highlighted poorer performance in entertainment-related content, suggesting that data scarcity in certain cultural and informal registers continues to hinder model generalisation.

Our findings underscore several broader issues in multilingual NLP: persistent infrastructural barriers for African languages, limited documentation and support in model repositories, and challenges in reproducibility due to computational costs. Moreover, the time-sensitive nature of benchmark corrections complicates longitudinal analysis and poses difficulties for future evaluation efforts.

Overall, this work contributes to the conversation around equitable NLP practices by critically assessing the robustness of current MT systems in low-resource settings. While automatic metrics show limited sensitivity to corrected references, our study reinforces the need for continuous improvement in both dataset quality and model inclusivity to advance the field toward more linguistically fair and socially responsible AI systems.

References

Abdulmumin, I., Mkhwanazi, S., Mbooi, M., Muhammad, S.H., Ahmad, I.S., Putini, N., Mathebula, M., Shingange, M., Gwadabe, T., & Marivate, V. (2024). *Correcting FLORES Evaluation Dataset for Four African Languages*. Proceedings of the Ninth Conference on Machine Translation.<https://doi.org/10.18653/v1/2024.wmt-1.44>American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Tiedemann, J. (2020). OPUS-MT: Building open translation services for the World. *arXiv preprint arXiv:2005.05943*.Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

Kudugunta, S. et al. (2023) ‘MADLAD-400: A Multilingual And Document-Level Large Audited Dataset’, arXiv [cs.CL] [Preprint]. Available at: <http://arxiv.org/abs/2309.04662>.

Team, Facebook AI. (2022). NLLB: No Language Left Behind.<https://ai.facebook.com/research/no-language-left-behind>

‘No Language Left Behind: Scaling Human-Centered Machine Translation’ (2022).

Figure 2

Figure 1

Figure 4

Figure 3