Impact of Reference Corrections on Machine Translation Model Rankings in Low-Resource African Languages

Authors:

Alec Watson u22491351, Aren Repko u04515791, Daniel Geerdink u22556860

Group 7

1. Introduction

The flores datasets for Hausa (hau), Northern Sotho (nso), Xitsonga (tso), isiZulu (zul) have all been corrected through manual intervention (Abdulmumin et al., 2024) and we would like to investigate the effect this has had on the quality of machine translations.

In particular, how do the dataset corrections impact the performance rankings of existing machine translation models and is the performance of these models highly dependent on model types or the domain of the text in question?

2. Background

Many MT systems for African languages are trained and evaluated using publicly available datasets like FLORES-200. However, several recent studies (Abdulmumin et al., 2024) have revealed serious errors in these test sets. Since automatic metrics like BLEU, COMET, and BERTScore rely on reference quality to judge translation accuracy, even small flaws in the reference can lead to unfair rankings or incorrect conclusions about system performance.

Furthermore, MT model outputs vary widely depending on model architecture (e.g., NLLB vs. OPUS-MT), training domain (e.g., religious text vs. news), tokenisation and vocabulary choices.

Understanding whether performance changes are due to better references or to model biases is key to building robust evaluations and fair comparisons.

3. Proposed Methodology

We will focus on four African languages in the FLORES (https://huggingface.co/datasets/openlanguagedata/flores_plus) benchmark and FLORES corrected data set(https://github.com/dsfsi/flores-fix-4-africa): Hausa, Xitsonga, isiZulu, and Northern Sotho. For each:

1. **Collect system outputs**: We will retrieve translation outputs from 3–5 publicly available MT models (e.g., NLLB, OPUS-MT, M2M-100).
2. **Score translations**: Using both the original and corrected references, we will score each model's outputs using BLEU, COMET and BERTScore.
3. **Analyse ranking shifts**:

    - Compare system rankings across original and corrected references.
    - Measure score changes per model per metric.
    - Calculate Spearman's rank correlation to quantify ranking consistency.
4. **Explore domain impact**:
    - Segment sentences by topic/domain (e.g., religion, news, government).
    - Analyse whether corrections have greater impact in some domains than others.

4. How we will measure success

We will evaluate success through:

- **Ranking Volatility**: If rankings differ significantly between original and corrected references, this quantifies the scale of the unreliability.
- **Score Delta**: Large differences in scores before/after corrections indicate metric sensitivity to reference quality.
- **Model Robustness**: If certain model types maintain stable rankings across datasets, this could show resilience to reference noise.
- **Domain Sensitivity**: If certain domains (e.g., technical or religious content) show larger shifts, this highlights where future benchmarks should focus.

5. Expected Outcomes and Contributions

We aim to deepen our understanding of how erroneous translations in FLORES affect machine translations in low resource African languages, considering factors like domain and model architecture, as discussed above. We anticipate difficulty in separating text into different domains across different languages. Going forward we hope this information will enable people to focus their correction efforts on whichever domain is most relevant to their ML usage and the architecture of their model.

References

- Abdulmumin, I., Mkhwanazi, S., Mbooi, M., Muhammad, S.H., Ahmad, I.S., Putini, N., Mathebula, M., Shingange, M., Gwadabe, T. and Marivate, V. (2024). Correcting FLORES Evaluation Dataset for Four African Languages. Proceedings of the Ninth Conference on Machine Translation, [online] pp.570–578. doi:https://doi.org/10.18653/v1/2024.wmt-1.44.