9th Conference on Machine Translation (WMT 2024), Miami, Florida, USA, 15 -16 November 2024

Correcting FLORES evaluation dataset for four African languages

Abdulmumin, I; Mkhwanazi, Sthembiso N; Mbooi, Mahlatse S; Muhammad, SH; Ahmad, IS; Putini, N; Mathebula, M; Shingange, M; Gwadabe, T; Marivate, V

Abstract

This paper describes the corrections made to the FLORES evaluation (dev and devtest) dataset for four African languages, namely Hausa, Northern Sotho (Sepedi), Xitsonga, and isiZulu. The original dataset, though groundbreaking in its coverage of low-resource languages, exhibited various inconsistencies and inaccuracies in the reviewed languages that could potentially hinder the integrity of the evaluation of downstream tasks in natural language processing (NLP), especially machine translation. Through a meticulous review process by native speakers, several corrections were identified and implemented, improving the overall quality and reliability of the dataset. For each language, we provide a concise summary of the errors encountered and corrected and also present some statistical analysis that measures the difference between the existing and corrected datasets. We believe that our corrections improve the linguistic accuracy and reliability of the data and, thereby, contribute to a more effective evaluation of NLP tasks involving the four African languages. Finally, we recommend that future translation efforts, particularly in low-resource languages, prioritize the active involvement of native speakers at every stage of the process to ensure linguistic accuracy and cultural relevance.