World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:15, No:04, 2021

Neural Machine Translation for Low-Resource African Languages: Benchmarking State-of-the-Art Transformer for Wolof

Authors : Cheikh Bamba Dione, Alla Lo, Elhadji Mamadou Nguer, Siley O. Ba

Abstract : In this paper, we propose two neural machine translation (NMT) systems (French-to-Wolof and Wolof-to-French) based on sequence-to-sequence with attention and transformer architectures. We trained our models on a parallel French-Wolof corpus of about 83k sentence pairs. Because of the low-resource setting, we experimented with advanced methods for handling data sparsity, including subword segmentation, back translation, and the copied corpus method. We evaluate the models using the BLEU score and find that transformer outperforms the classic seq2seq model in all settings, in addition to being less sensitive to noise. In general, the best scores are achieved when training the models on word-level-based units. For subword-level models, using back translation proves to be slightly beneficial in low-resource (WO) to high-resource (FR) language translation for the transformer (but not for the seq2seq) models. A slight improvement can also be observed when injecting copied monolingual text in the target language. Moreover, combining the copied method data with back translation leads to a substantial improvement of the translation quality.

Keywords : backtranslation, low-resource language, neural machine translation, sequence-to-sequence, transformer, Wolof
Conference Title : ICCLMT 2021 : International Conference on Computational Linguistics and Machine Translation
Conference Location : Tokyo, Japan
Conference Dates : April 22-23, 2021

Open Science Index, Cognitive and Language Sciences Vol:15, No:04, 2021 publications.waset.org/abstracts/135110.pdf