5th Conference on Information Communications Technology and Society, Durban, South Africa, 10 -11 March 2021

Transformer-based machine translation for low-resourced languages embedded with language identification

Tshephisho J. Sefara
Next Generation Enterprises and Institutions, Council for Scientific and Industrial Research
Pretoria, South Africa, tsefara@csir.co.za

Nelisiwe Gama
School of Computer Science and Applied Mathematics, University of the Witwatersrand
South Africa, nellygrattie@gmail.com

Phillemon N. Senoamadi
Department of Mathematics, University of Zululand, South Africa, phillemon@aims.ac.za

Skhumbuzo G. Zwane
Department of Computer Science, University of Zululand, South Africa,
201144122@stu.unizulu.ac.za

Hlawulani Sibisi
Department of Computer Science, University of Johannesburg, South Africa,
hlawusibisi@gmail.com

Vukosi Marivate
Department of Computer Science, University of Pretoria, South Africa,
vukosi.marivate@cs.up.ac.za

https://ieeexplore.ieee.org/document/9394996

Abstract

Recent research on the development of machine translation (MT) models has resulted in state-of-the-art performance for many resourced European languages. However, there has been a little focus on applying these MT services to low-resourced languages. This paper presents the development of neural machine translation (NMT) for lowresourced languages of South Africa. Two MT models, JoeyNMT and transformer NMT with self-attention are trained and evaluated using BLEU score. The transformer NMT with self-attention obtained state-of-the-art performance on isiNdebele, SiSwati, Setswana, Tshivenda, isiXhosa, and Sepedi while JoeyNMT performed well on isiZulu. The MT models are embedded with language identification (LID) model that presets the language for translation models. The LID models are trained using logistic regression and multinomial naive Bayes (MNB). MNB classifier obtained an accuracy of 99% outperforming logistic regression which obtained the lowest accuracy of 97%.