A Comparison of Data Augmentation Techniques for Nguni Language Statistical and Neural Machine Translation models

Machine translation(MT) refers to the use of computer systems to perform translations between languages. This field has seen great advances in recent times aided by the advent of the internet and neural machine translation. However, the translation of Nguni languages cannot boast the same rise quality. This is owed to the limited amount of training data available for these languages. As such our project aimed to compare data augmentation techniques for Nguni language machine translation by comparing MT models trained on data augmented with synthetic data generated using back-translation, MT models trained on multilingual data and baseline MT models trained only with the original available data for English to isiZulu and English to isiXhosa translation. Each of these 3 types of systems were implemented using a neural machine translation system and a statistical machine translation system. The quality of each system was evaluated using the BLEU technique which yields a score between 0 and 100, 100 being a perfect translation.

Research Questions

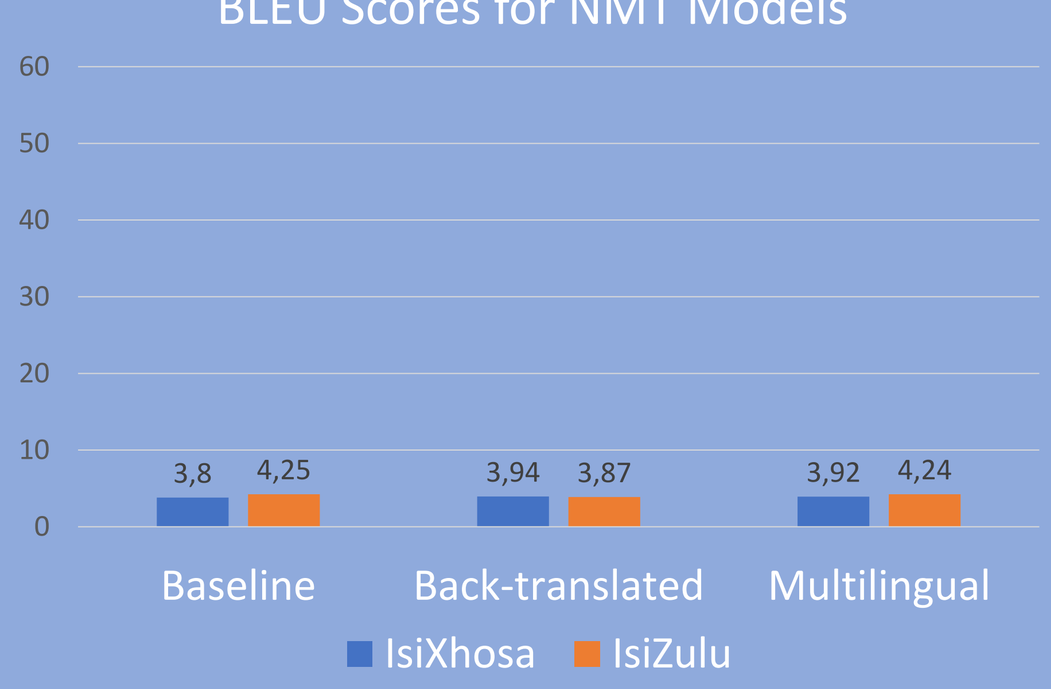
1. Do MT systems trained on data augmented using backtranslation have higher BLEU scores for Nguni translations?
2. Do MT systems trained on multilingual data have higher BLEU scores for Nguni translations?

Hypothesis

For both SMT and NMT systems, and both isiXhosa and IsiZulu translations, the machine translation models trained using multilingual data will yield the highest BLEU scores, followed by the back-translation models, with the baseline models yielding the lowest BLEU scores.

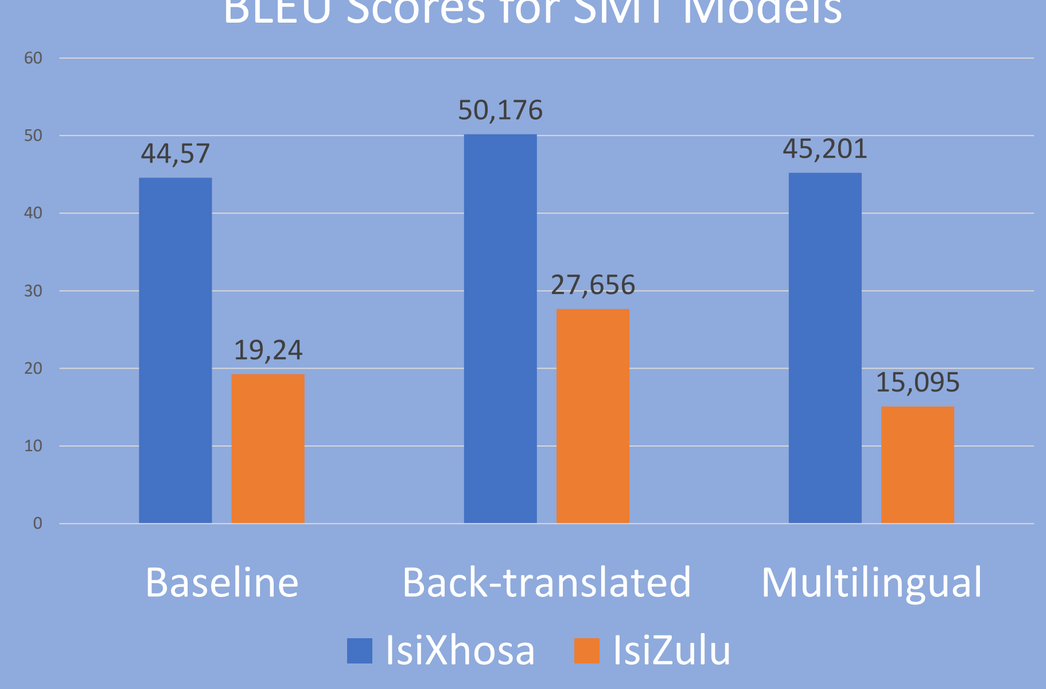
NMT Systems Performance

BLEU Scores for NMT Models



SMT Systems Performance

BLEU Scores for SMT Models



Conclusions

In this study we trained transformer neural machine translation (NMT) systems and statistical machine translation (SMT) systems on data augmented with back-translated data and multilingual data and compared these against baseline MT systems. In the NMT context, both the multilingual and back-translated systems outperformed the baseline systems for English-to-isiXhosa translation, a similar conclusion was reached in the SMT context. For English-to-isiZulu translation however, in the SMT context the baseline systems outperformed the multilingual with the back-translation system yielding the best BLEU scores for both isiZulu and isiXhosa translation. In the NMT context, the baseline model for IsiZulu was seen to overfit the data due to its small size, thus, the baseline system had the best performance, with the back-translated systems outperforming the multilingual models in both isiXhosa and isiZulu cases.

SCHOOL OF IT

University of Cape Town
Department of Computer Science
Email: dept@cs.uct.ac.za
Tel: 021 650 2663

Deevesh Beegun
deeveshbeegun@gmail.com
Fezeka Nzama
nzmfez001@myuct.ac.za
Supervised by Jan Buys