ISIXHOSA MEDICAL MACHINE TRANSLATION

1.Introduction

IsiXhosa is South Africa's second-most spoken home language, yet many medical services operate primarily in English. This language barrier creates serious health risks through misdiagnoses and treatment errors. Machine translation (MT) could bridge this gap, but isiXhosa medical translation faces a critical obstacle: scarcity of parallel medical training data for isiXhosa. Synthetic data generation offers a potential solution to overcome this limitation.

2. Objective

- To investigate whether synthetic data generation through back-translation and DALI improves medical English-isiXhosa translation quality
- Compare training models from scratch versus fine-tuning existing multilingual models
- Assess performance using standard MT metrics (BLEU, chrF and chrF++), and a health term error rate.

3. Methodology

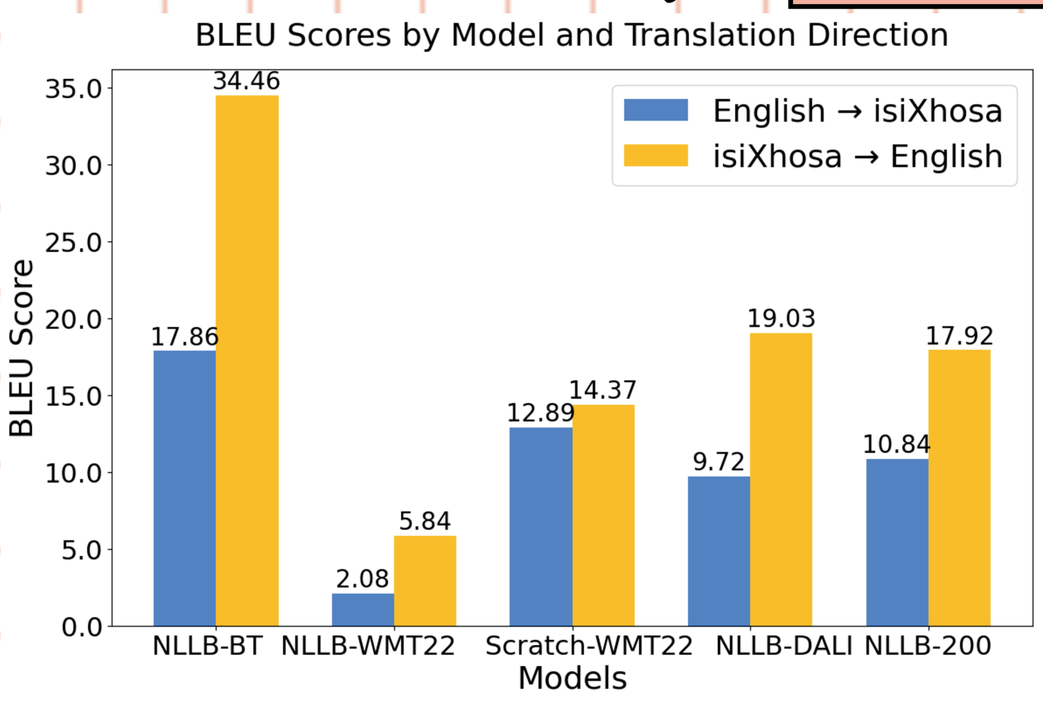
Data Generation

- Use PriMock47 as the English source text for generating the synthetic data
- Back-translate PriMock47 using NLLB-200 to get synthetic isiXhosa
- Use DALI which uses a bilingual lexicon to translate PriMock47 into synthetic isiXhosa

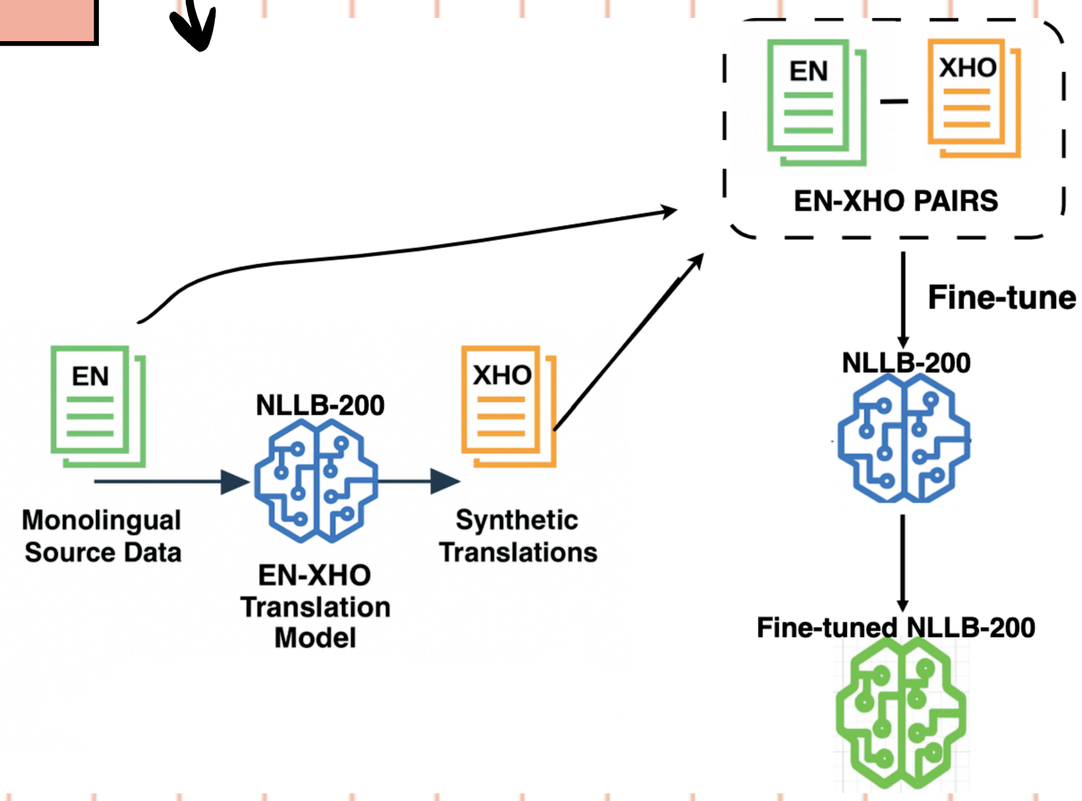
Training

- Baseline models:
  - Use Fairseq to train a transformer model on the English-isiXhosa subset of WMT22
  - Fine-tune NLLB-200 on WMT22
- Synthetic data models, fine-tune NLLB-200 on:
  - Back-translated PriMock47
  - DALI generated PriMock47

4. Results



5. NLLB-BT Method



6. Analysis

- **NLLB-BT** performed best showing significant BLEU gains over the baseline, particularly strong for isiXhosa→English
- **NLLB-DALI** degraded performance likely due to poor lexicon quality leading to low-quality synthetic data, underperforming compared to the baseline
- **Scratch-WMT22** beat **NLLB-WMT22** showing that from-scratch training outperformed fine-tuning
- Combining real data with back-translated synthetic data outperformed pure synthetic approaches

7. Conclusions

- When parallel corpora are scarce, back-translation proves to be an effective synthetic data generation strategy
- Domain adaptation improved both medical and general-domain translation for **NLLB-BT**
- Best models were produced by mixing real and synthetic data
- Future work includes experimenting with larger medical source corpora

Team Members

Nick Matzopoulos	mtznic006@myuct.ac.za
Malibongwe Makhonza	mkhmal024@myuct.ac.za
Elijah Sherman	shreli006@myuct.ac.za

Supervisor

Francois Meyer	francois.meyer@uct.ac.za