Exploring Machine Translation for code-switching between English and Setswana in South African classrooms

DEEP LEARNING INDABA X Uganda

Keneilwe Mokoka

Research Aim

- The aim of this research is to leverage existing PLMs such as the mT5 and M2M-100 to translate mathematics text from English to English/Setswana as an attempt to aid the Department of Education in South Africa to solve the challenge of low numeracy skills amongst learners.

Research Objectives

- To collect mathematical English text and corresponding code-switched English/Setswana translation.
- To build a code-switched corpus for training.
- To fine-tune the MT5 model and evaluate its performance on translating mathematical English to English-Setswana mathematical text.
- To fine-tune the M2M-100 model and evaluate its performance on translating mathematical English to English-Setswana mathematical text

References

[1] Adam Siepel, Katherine S Pollard, and David Haussler. New methods for detecting lineage-specific selection. In *Research in Computational Molecular Biology*, pages 190–205. Springer, 2006.

[2] Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. The meme suite. *Nucleic acids research*, page gkv416, 2015.

[3] Heila Jordaan. Language teaching is no panacea: A theoretical perspective and critical evaluation of language in education within the south african context. *South African Journal of Communication Disorders*, 58(2):79, 2011.

[4] Mmaki Jantjies and Mike Joy. Multilingual mobile learning: a case study of four south african high schools. In *Proceedings of the 11th International Conference on Mobile and Contextual Learning 2012*, pages 208–211. CEUR-WS, 2012.

[5] Nikki Stein. Language in schools. *Basic education rights handbook: Education rights in South Africa. Johannesburg, South Africa: SECTION27. Available at https://section27. org. za/wpcontent/uploads/2017/02*, 2017.

[6] Sally-Ann Robertson and Mellony Graven. Language as an including or excluding factor in mathematics teaching and learning. *Mathematics Education Research Journal*, 32(1):77–101, 2020.

[7] Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *arXiv preprint arXiv:2212.09660*, 2022.

[8] Ramaesela Jerminah Khobo. *The effect of using computers for the teaching and learning of mathematics to grade 10 learners at secondary school.* PhD thesis, University of South Africa, 2015.

[9] Richard Lastrucci, Isheanesu Dzingirai, Jenalea Rajab, Andani Madodonga, Matimba Shingange, Daniel Njini, and Vukosi Marivate. Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora. *arXiv preprint arXiv:2303.03750*, 2023.

[10] Ronny Mabokela and Tim Schlippe. A sentiment corpus for south african under-resourced languages in a multilingual context. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 70–77, 2022.

Introduction

- A major challenge that the Department of Education (DoE) in South Africa (SA) is facing is the low achievement levels of numeracy and literacy amongst most school children [3, 4].
- Research attributes this challenge to an inadequate proficiency in the language of learning and teaching (LoLT), amongst other factors [3, 5].
- Research has shown that switching (also referred to as code-switching or CSW) between English as a 2nd language (L2) and the learners' 1st language (L1) can greatly improve the learners' understanding of Mathematics [6] in a multilingual society like SA [4].
- CSW has been a topic of interest in the Natural Language Processing (NLP) research space in the past decade, as a result of the majority of the population in the world being multilingual [7].
- This has inspired the focus of this research: translating mathematical text from English to English mixed with Setswana.

Research Design

- This research will be done in a school setting. It will focus on collecting data in the form of audio recording of mathematical lessons given to Grade 10 learners.
- The interest in grade 10 learners stems from the fact that in SAn schools, learners are given the opportunity to choose their subjects of interest in grade 9 for the remainder of their high school years. [8].

Data

- The data collection will be done in the form of audio recordings of mathematical lessons given to grade 10 learners of Thethe High School based in Luka which is a village outside of Rustenburg.
- It is envisaged that one week's worth of recording will be sufficient, with the flexibility of more time if necessary. Each lesson has a duration of 3 hours and a total of 15 hours worth of audio will be obtained.
- The data will be transcribed and filtered accordingly to build the corpus.
- To ensure data accuracy, integrity and credibility, a mathematical teacher will be requested to review the manually curated corpus, to ensure that the translations are correct.

Data Pre-Processing

The recordings will be manually transcripted. The data pre-processing is expected to follow the following steps for the creation of a parallel corpus:

- English sentences with their corresponding code-switched translations will be identified and split into source and target text and placed into a text file[9].
- Any duplicate texts will be removed [9].
- All punctuation will be removed [9, 10].
- All text will be converted to lower case [9, 10].
- For alignment, the text will be taken through the NLTK Tokenizer: punkt, to get the vector of the sentence tokens [9].
- Following the guidelines of [9], the n tokenized sentences will be encoded to n sentence vectors through the use of Language-Agnostic Sentence Representations (LASER) which is a research project by Facebook AI Research. The cosine similarities of the vectors will be calculated and will be included in the corpus [9].

Ethical Considerations

- Permission to record the lessons has been requested.
- Consent has been received from the educator and the learners' parents/legal guardians since they are minors.
- Awaiting ethical clearance from the School of Computer Science and Applied Mathematics before collecting data.
- The data extracted for the corpus will exclude names of people and places. It is only concerned with mathematical concepts.
- While the corpus will be made available to the general public for further NLP research, the audio recordings will only be accessible to the researcher.