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

Keneilwe Mokoka*
School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg
Deep Learning Indaba Ghana
Accra, Ghana
keneilwe.mafatshe@gmail.com

Abstract

One of the major challenges of the Department of Education in South Africa is the low numeracy skills amongst South African learners. This study seeks to spotlight the low numeracy challenge encountered in the learning and teaching of mathematics in a classroom where the educator and learners are Setswana native speakers, but use English as the language of learning and teaching. Using English as a language of learning and teaching mathematics to non-native English speaking learners has been stated as one of the reasons why learners perform poorly in mathematics, leading to low numeracy skills. It has been shown that when educators code-switch between English and the native language of the learners to explain mathematical concepts, learners tend to participate more in the classroom and perform better in mathematics. Codeswitching is a topic of interest in Natural Language Processing. Pretrained language models (PLM), such as the mT5, have previously been used in machine translation of code-switched data. In this research, a small corpus of parallel text consisting of monolingual English mathematical text translated into English-Setswana (code-mixed English and Setswana) will be used to fine-tune the mT5 PLM for the purpose of translating mathematical text from English to English/Setswana. In addition, the M2M-100 PLM has been leveraged by African researchers for the machine translation of low-resourced languages. This research aims to fine-tune the M2M-100 PLM using the same corpus and to evaluate its performance on the task of machine translation to potentially aid in the learning and teaching of mathematics in South Africa.

1 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 Jordaan [2011], Jantjies and Joy [2012]. Research attributes this challenge to a number of factors which include teachers' lack of subject matter understanding, poverty, absenteeism, learners not being on time for school resulting in missing instructional time as well as inadequate proficiency in the language of learning and teaching (LoLT) Jordaan [2011], Stein [2017]. Special attention has been given to the subject of Mathematics because competency in this subject does not only contribute to the improved quality of life of the individual through future work opportunities as well as studying further, but is also essential to improving the economic status of a society Robertson and Graven [2020].

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

5th Deep Learning Indaba Conference (DLI 2023).

SA has 11 official languages, yet the English language is predominantly used as the LoLT in most schools for various reasons (which are not confined to SA) such as the effects of colonialism and the perception of English being the language of economic opportunity Jordaan [2011], Setati [2001]. Research has shown that learners that are taught Mathematics with a second language (L2) as the medium of instruction tend to perform poorer than their counterparts that are taught in their first language (L1) Jordaan [2011]. Many South African (SAn) children are multilingual as a result of exposure to multiple languages at home and within the community, in addition to the English language Jordaan [2011].
Most African languages cannot be used to teach subjects such as Mathematics since their vocabulary has not yet been developed for such subjectsSetati [2001]. For that reason research has shown that switching (also referred to as code-switching or CSW) between English as a L2 and the learners' L1 can greatly improve the learners' understanding of Mathematics Robertson and Graven [2020] in a multilingual society like SA Jantjies and Joy [2012].
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 Winata et al. [2022]. It has been applied to areas such as sentiment analysis, parts of speech (PoS) tagging and machine translation (MT), amongst others Winata et al. [2022]. In their study, Jaharah et al (2021) used NLP models to translate English text into Hindi mixed with English. Aguilar and Solorio (2019) adapted models that are monolingual to code-switched tasks such as language identification, for language pairs such as Nepali-English. Mabokela and Schlippe (2021) conducted one of the few NLP researches that focused on low resource African languages. This has inspired the focus of this research: translating mathematical text from English to English mixed with Setswana.

1.1 Literature Review

1.1.1 Language in mathematical education in SA

Passing Mathematics is a pre-requisite to pursuing a qualification in fields with high-income potential, such as engineering and medicine, amongst others, in higher institutions of learning in SA Setati [2001]. Yet, research has shown that 61% of grade 5 learners in SA cannot perform basic mathematical operations such as addition, subtraction, and multiplication Robertson and Graven [2020]. Factors such as poverty, absenteeism and inadequate proficiency in the LoLT have been highlighted as potential reasons of the low levels of numeracy skills amongst SAn learners Jordaan [2011], Robertson and Graven [2020].
Language plays a major role in expressing, developing, and thinking about mathematics Setati [2001]. English is used as the preferred LoLT for 80% of SAn learners, yet it is L1 for less than 10% of the country's population Robertson and Graven [2020]. Research has shown that learners that are taught and assessed in their native language perform better than those that are taught and assessed in their non-native language Setati [2001]. At a school based in the Eastern Cape province of SA where English is used as the LoLT and with learners with isiXhosa as their L1 , Robertson and Graven (2020) observed that the mathematics teacher opted for tasks that were not cognitively demanding, as a result of their learners' linguistics limitations Robertson and Graven [2020]. It was further observed that teachers do not make much of an effort to compel learners to be inquisitive and have discussions around the reasons for certain mathematical concepts due to learners' linguistic limitations Robertson and Graven [2020]. This in turn prevents learners from fully understanding the concepts of mathematics Robertson and Graven [2020].
There is also the challenge of African languages not being developed for subjects such as mathematics, therefore these languages cannot solely be used by teachers to teach mathematics Setati [2001]. To overcome this hurdle, researchers support multilingual teachers in using switching between English and the learners' L1 in teaching mathematics Gaoshubelwe [2011], Robertson and Graven [2020], Setati [1998]. In their paper, Gaoshubelwe (2011) suggests the use of CSW as an additional teaching resource for teaching mathematics, to encourage learners to have more exploratory discussions, while they are in the process of improving their proficiency in the English language. This follows the observation of three teachers from three different schools that use English as a LoLT, who relied on code-switching between English and Setswana to teach mathematics to learners with Setswana as their main language Gaoshubelwe [2011]. These teachers mainly used CSW to help decode mathematical terminology, so that learners can better conceptualize mathematics Gaoshubelwe [2011].
In a study, Setati (1998) observed that the mathematics teacher code switched between Setswana and English for the purpose of reformulating concepts in the form of paraphrasing, to explain concepts

and to regulate the classroom as well as to directly translate terms. The researcher observed that learners participated in lessons more when the teacher encouraged CSW, as they felt comfortable to communicate using their main language which is Setswana Setati [1998]. CSW is also being used as a tool to help clarify concepts of subjects as well as to bring students and teachers closer in schools in the country of Botswana Chimbganda and Mokgwathi [2012].

1.1.2 Code-Switching and NLP

CSW is defined as a scenario where a multilingual speaker utilises more than one language as they make utterances in a single conversation Nguyen et al. [2022], Winata et al. [2022]. In the case of NLP, CSW has been popularly applied in language detection, sentiment analysis, named entity recognition (NER), PoS tagging, offensive language detection, automatic speech recognition (ASR), language modeling and MT topics Winata et al. [2022]. For their research, Jawahar et al (2021) explored using text-to-text transformers for the machine translation of text from English to Hinglish (English mixed with Hindi). This MT system achieved this by transforming a monolingual sequence of words into a code-mixed sequence of words Jawahar et al. [2021]. For the NLP task of language identification, Aguilar and Solorio (2019) adapted NLP models that have been previously trained on English to identify languages in data that has been code-switched for multiple language pairs.

In an African context, Mabokela and Schlippe (2022) presented a sentiment corpus for three SAn spoken languages: English, Sepedi, and Setswana as an attempt to include under-resourced languages in NLP research. This multilingual corpus, containing CSW, was used for sentiment analysis to detect social challenges in SA Mabokela and Schlippe [2022a]. There is potential to build educational technologies to cater to code switching, Nguyen et al (2022) noted that the existing educational platforms were not designed with methods of teaching in mind. Nguyen et al (2022) also noted that technologies that were designed for the purpose of teaching were done so with a monolingual bias. One of the reasons for these constraints is the lack of data Nguyen et al. [2022].

NLP has also been applied in CSW by using statistical methods such as Naive Bayes and support vector machine (SVM) for text classification and conditional random field (CRF) for PoS tagging Winata et al. [2022]. Neural based methods seem to be more popular than statistical methods for CSW research, Recurrent Neural Networks (RNN) and long-short memory (LSTM) architectures have been used for tasks such as language modeling and CSW identification, where DNN-based and hybrid HMM-DNN models are used in speech recognition models Winata et al. [2022].

1.1.3 Overcoming low-resourced languages for NLP research challenges

A major challenge in CSW in NLP research is the availability and accessibility of data (in cases where there is data available) and it is reported in Winata et al (2022) that most CSW in NLP research has been done for Spanish-English, Modern Standard Arabic-Egyptian and Hindi-English language pairs. It has been reported that most researchers source data for their research from social media Winata et al. [2022], as is the case with Mabokela and Schlippe [2022b] and Jawahar et al. [2021]. Speech data (recordings of conversations and interviews), speech transcription, news, dialogues, books, government documents, and treebanks are reported to also be common sources of data for CSW in NLP research Winata et al. [2022]. For example, to build their multilingual corpus, Lastrucci et al 2023) sourced their data from a magazine that is published by the SAn government.

There is a lack of representation of African languages in NLP research, even though African languages constitute 30.15% of languages that are in use around the world today Martinus et al. [2020]. Africans are working hard to bridge the gap Martinus et al. [2020]. Data collection is not always feasible for low resourced languages since these languages lack representation on the web Adelani et al. [2022]. To get over this hurdle, some researchers have opted for synthetic data as noted by Winata et al (2022) by using generative adversarial networks (GANs) to generate CSW sentences. The generator in turn learns to predict CSW points without needing to know the language Winata et al. [2022]. Jawahar et al (2021) generated their own code-mixed data to assist with fine-tuning their models.

Mehnaz et al (2021) presented an annotated dataset consisting of code-switched conversations to be used for the NLP task of abstractive summarization Mehnaz et al. [2021]. In their research Adelani et al (2022), achieved satisfactory results when fine-tuning pre-trained language models with limited data to potentially alleviate the challenge of the lack of data for low resourced African languages in NLP research Adelani et al. [2022].

1.1.4 Leveraging Pre-trained Language Models for Machine Translation

Transfer learning is commonly used in NLP pipelines, where a language model that was pre-trained on a task that is rich in data is fine-tuned for a downstream NLP task Xue et al. [2020]. For example, Jawahar et al (2021) utilized the language model mT5 which was pre-trained on a multilingual dataset consisting of over 100 languages to translate English text into English-Hinglish text.
As an attempt to include more languages for MT, Fan et al (2021) created a non-English centric Many-to-many multilingual Pre-trained Language Model (PLM) on 100 languages which they refer to as M2M-100 Fan et al. [2021]. This model has the ability of directly translating between any pair of 100 languages Fan et al. [2021]. In their research, Adelani et al (2022) fine-tuned the M2M-100 PLM for multilingual translations of African languages to be used for data that is only available in small quantities, which is common for low-resourced languages such as most African languages Adelani et al. [2022]. Lastrucci et al (2023) also fine-tuned the M2M-100 PLM to build baseline translation benchmarks for their Vuk'uzenzele and ZAgov-multilingual datasets Lastrucci et al. [2023]. The following sections provide brief descriptions of the mT5 and M2M-100 PLMs.

mT5 : A multilingual text-to-text transformer mT5 is a PLM that was built to aid in transfer learning for NLP research Xue et al. [2020]. The creation of mT5 was inspired by its predecessor T5, which is a text-to-text transformer, that was meant for NLP tasks that focus on the English language Xue et al. [2020]. The "text-to-text" refers to the concept of a model using text as input, and generates text as output Raffel et al. [2020]. Both these models are based on the encoder-decoder transformer architecture Raffel et al. [2020]. The transformer architecture performs well on the task of MT, and has subsequently been used for other NLP tasks Vaswani et al. [2017], Raffel et al. [2020]. mT5 was trained on a dataset referred to as mC4 that consists of languages including some of our African languages such as Afrikaans, Shona, Sotho, Swahili, Xhosa, and Zulu, where T5 was pre-trained on a dataset (C4 aka "Colossal Cleaned Crawled Corpus") that consists of just the English languageXue et al. [2020]. Both these datasets were sourced from the public common crawl web scrape, which is a repository of data (web corpus) collected from the web that can be used for NLP research Luccioni and Viviano [2021].

M2M-100 : A Many-to-Many translation model The M2M-100 PLM was created for the purpose of machine translation between any set of 100 languages Fan et al. [2021]. This was done to cater to more translation needs across other languages than the English language Adelani et al. [2022]. This PLM follows the transformer sequence to sequence architecture which comprises of the encoder and decoder Fan et al. [2021]. The dataset that was used for training has various sources which include the workshop on machine translation, the workshop on Asian translation, the international conference on spoken language translation, FLORES, TED, Autshumato, and Tatoeba Fan et al. [2021].

2 Research Question

The research question is as follows:

How do the Pre-trained Language Models mT5 and M2M-100 fare in the task of translating mathematical text from English to English mixed with Setswana?

3 Research Aim and Objectives

3.1 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.

3.2 Research Objectives

The objectives of this research are:

- 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

4 Research Methodology

The primary focus of this study is to build a multilingual corpus consisting of English mathematical text and code-switched English/Setswana mathematical text and run it through the PLM MT5 and M2M-100 models to explore the possibility of NLP aiding in the teaching and learning of mathematics in South African multilingual classrooms.

5 Conclusion

Educators use English as a LoLT to teach mathematics in SAn classrooms where the majority of learners are non-native English speakers Setati [1998]. This is reported as one of the reasons why SAn schools face low achievement levels of numeracy amongst learners Jordaan [2011]. Educators then resort to code-switching between English and the learners' native language Jordaan [2011], to aid in explaining concepts and encouraging learner participation Setati [1998]. In a mathematical classroom of learners that are native Setswana speakers, it was observed that when the educator mixed languages (English and Setswana) during lessons, learners were more willing to participate and performed better in mathematics Setati [1998]. Code-switching is not only limited to the classrooms, Jawahar et al. (2021) as well as Mabokela and Schlippe (2022) have noted that social media users utilize code switching all the time Jawahar et al. [2021], Mabokela and Schlippe [2022b]. Because of that, there has been interest in NLP research for CSW Jawahar et al. [2021]. 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 potentially aid the Department of Education to solve the challenge of low numeracy skills amongst learners in South Africa.

References

D. I. Adelani, J. O. Alabi, A. Fan, J. Kreutzer, X. Shen, M. Reid, D. Ruiter, D. Klakow, P. Nabende, E. Chang, et al. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *arXiv preprint arXiv:2205.02022*, 2022.

A. B. Chimbganda and T. S. Mokgwathi. Code-switching in botswana's esl classrooms: A paradox of linguistic policy in education. *International Journal of English Linguistics*, 2(2):21, 2012.

A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886, 2021.

B. S. Gaoshubelwe. *Language practices in the teaching and learning of mathematics: A case study of three mathematics teachers in multilingual schools*. PhD thesis, North-West University, 2011.

M. Jantjies and M. 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.

G. Jawahar, E. M. B. Nagoudi, M. Abdul-Mageed, and L. V. Lakshmanan. Exploring text-to-text transformers for english to hinglish machine translation with synthetic code-mixing. *arXiv preprint arXiv:2105.08807*, 2021.

H. 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.

R. Lastrucci, I. Dzingirai, J. Rajab, A. Madodonga, M. Shingange, D. Njini, and V. Marivate. Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora. *arXiv preprint arXiv:2303.03750*, 2023.

A. Luccioni and J. Viviano. What's in the box? an analysis of undesirable content in the common crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, 2021.

R. Mabokela and T. Schlippe. Ai for social good: Sentiment analysis to detect social challenges in south africa. In *Artificial Intelligence Research: Third Southern African Conference, SACAIR 2022, Stellenbosch, South Africa, December 5–9, 2022, Proceedings*, pages 309–322. Springer, 2022a.

R. Mabokela and T. 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, 2022b.

L. Martinus, J. T. Ali, J. Abbott, V. Marivate, S. Kabongo, M. Meressa, E. Murhabazi, O. Ahia, E. van Biljon, A. Ramkilowan, et al. Masakhane-machine translation for africa. *arXiv preprint arXiv:2003.11529*, 2020.

L. Mehnaz, D. Mahata, R. Gosangi, U. S. Gunturi, R. Jain, G. Gupta, A. Kumar, I. Lee, A. Acharya, and R. R. Shah. Gupshup: An annotated corpus for abstractive summarization of open-domain code-switched conversations. *arXiv preprint arXiv:2104.08578*, 2021.

L. Nguyen, Z. Yuan, and G. Seed. Building educational technologies for code-switching: Current practices, difficulties and future directions. *Languages*, 7(3):220, 2022.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

S.-A. Robertson and M. Graven. Language as an including or excluding factor in mathematics teaching and learning. *Mathematics Education Research Journal*, 32(1):77–101, 2020.

M. Setati. Code-switching in a senior primary class of second-language mathematics learners. *For the Learning of Mathematics*, 18(1):34–40, 1998.

M. Setati. Researching mathematics education and language in multilingual south africa. *The mathematics educator*, 12(2), 2001.

N. 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.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

G. I. Winata, A. F. Aji, Z.-X. Yong, and T. Solorio. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *arXiv preprint arXiv:2212.09660*, 2022.

L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.