The Analysis of the Sepedi-English Code-switched Radio News Corpus

Ramalepe, Simon
Computer Science Department, University of Limpopo, Sovenga, 0727, South Africa.
simon.ramalepe@ul.ac.za

Modipa, Thipe I.
Computer Science Department, University of Limpopo, Sovenga, 0727, South Africa.
Centre for Artificial Intelligence Research (CAIR), South Africa.
thipe.modipa@ul.ac.za

Davel, Marelie H.
Faculty of Engineering, North-West University, Potchefstroom, South Africa
Centre for Artificial Intelligence Research (CAIR), South Africa.

Abstract

Code-switching is a phenomenon that occurs mostly in multilingual countries where multilingual speakers often switch between languages in their conversations. The unavailability of large-scale code-switched corpora hampers the development and training of language models for the generation of code-switched text. In this study, we explore the initial phase of collecting and creating Sepedi-English code-switched corpus for generating synthetic news. Radio news and the frequency of code-switching on read news were considered and analysed. We developed and trained a Transformer-based language model using the collected code-switched dataset. We observed that the frequency of code-switched data in the dataset was very low at 1.1%. We complemented our dataset with the news headlines dataset to create a new dataset. Although the frequency was still low, the model obtained the optimal loss rate of 2,361 with an accuracy of 66%.

Keywords: Code-switching, text generation, radio news, Transformers, Sepedi

1 Introduction

Code-switching is the use of more than one language within a sentence in a discourse. It is generally more prevalent in multilingual communities through speech than text (Gupta et al. 2020). However, the current trend of communication through technology involves text, and it is more eminent on social media. Like in most multilingual countries, South Africans, notably Sepedi speakers use code-switching in their conversations. Code-switching is generally categorised into inter-sentential and intra-sentential switching. Inter-sentential code-switching occurs on a sentence level when a speaker switches languages from one sentence to another while intra-sentential occurs on a word level when more than one language is used in a sentence (Hamed et al. 2017). The primary language in a code-switching environment is commonly known as the matrix language while secondary language is known as the embedded language (Hamed et al. 2017).

Another form of intra-sentential code-switching uses borrowed or loanwords (Chan et al. 2005). Borrowed or loaned words occur in situations where vowels or consonants are either added or replaced to reproduce phonetically accepted words in the matrix language. The distinction between code-switching, code-mixing, and borrowing is difficult hence, in this study we will use code-switching to refer to both inter and intra-sentential code-switching.

Studies have shown that there is sparsity or lack of code-switched text data (Chang et al. 2018, Gao et al. 2019, Gupta et al. 2020). This could be a result of the informal nature of code-switched data. In other words, code-switching normally occurs on an informal or social environment either in speech or text. Hence, there is a lack of documented datasets on code-switched data. Development of synthetically generated code-switched corpus that can be used to train language models for text generation has been proposed. However, the generated text is not au-

thentic and has to be tested if it resembles real code-switched phenomenon. The lack of existing large-scale code-switched corpora is a challenge to the development of code-switched text generation language models, especially for under-resourced languages such as Sepedi.

Sepedi is one of the official languages of South Africa and is classified as an under-resourced language. Currently, the available code-switched corpora for the Sepedi-English language pair is not enough for use with deep learning techniques. In this study, we discuss and analyse the initial process of collecting Sepedi-English code-switched text data for the development of code-switched models for this language pair. The collected corpus is then applied to a text generation language model to observe the performance of the model using non synthetically generated code-switched text data.

The paper is structured as follows: Section 2 discusses the background of the study. Section 3 discusses the datasets used for training the text generation model. In Section 4 we discuss the experiments conducted, while in Section 5 we focus on the results obtained and the evaluation of the model thereof. Concluding remarks and future work are made in Section 6.

2 Background

Sepedi language is mainly spoken in the Limpopo province of South Africa with a population of 5.9 million (Statistics South Africa 2022). It is the matrix language in most spontaneous conversations in the province while English and other South African languages become embedded in the conversations. Code-switching is not common in read speech or written text but occurs quite often in spontaneous conversations which makes it difficult to collect code-switched data for training and testing language models. When it does occur it is for emphasis in a multilingual setting. Attempts to collect code-switched speech corpora have been made. Hamed et al. (2018), collected spontaneous Egyptian Arabic-English code-switched speech data by conducting and recording informal interview conversations. Analysis of their data showed that there was high usage of code-mixing (intra-sentential). Of the 1,234 sentences in the their dataset, 985 (79.8%) sentences were code-mixed, 124 (10%) sentences were Arabic monolingual text and 125 (10.1%) sentences were English monolingual text. The most frequently trigger words preceding the code-switching point were also noted along with the most frequent uni-grams, bi-grams, and tri-grams. Part-of-speech (POS) tagging was done to a portion of data and it was noted that nouns are used mostly in the embedded language.

In another study, Chan et al. (2005) developed a Cantonese-English code-mixing speech corpus to study the effect of Cantonese accent in English. A situation where most of the English words contain a Cantonese accent. For data collection, newsgroups and online diary methods were used. The frequency of code-switched words, part-of-speech tagging of the code-mixed words, and the length of the code-switched text were noted. Like in Hamed et al. (2017), the frequency of noun occurrence was high at 62.3% and words with length 1 were 74.96% Lyu et al. (2015) also, developed a Mandarin-English Code-switching Speech Corpus. Their corpus was dominated by intra-sentential code-switching of the matrix language. The data collection approach also included recorded interviews and conversations. The most frequent words were also identified.

Modipa et al. (2013), developed a Sepedi-English code-switched speech corpus to analyse the implication factors of code-switching when developing an automatic speech recognition (ASR) systems that are capable of dealing with Sepedi-English code-switched speech. In their data collection approach, radio broadcasts were recorded and the number of code-switched events was counted and transcribed to create the Sepedi prompted code-switched corpus (SPCS) (Modipa & Davel 2022).

Table 1: The number of sentences, words and unique tokens in the datasets.

Dataset	Categories	Total	Train/Val/Test
Radio News	#Sentences	501	350/100.1/50
	#Tokens	14 516	
	#Unique tokens	1 654	
	%English words	1.1%	
News Headlines	#Sentences	1 182	827/236/119
	#Tokens	16 135	
	#Unique tokens	2 781	
	% English words	25%	
Combined dataset	#Sentences	1 683	1 178/337/168
	#Tokens	30 654	
	#Unique tokens	3 824	
	% English words	26%	

Table 2: Top 10 most frequent English and borrowed words in the dataset.

English Word	**Occurrence**	**Borrowed Word**	**Occurrence**
National	11	praevete	1
Economic	12	Magistrata	1
Andrew	13	yaSouth	1
Congress	14	Peresente	1
Africa	14	Uniti	1
Eskom	15	Ekonomi	1
Democratic	17	Dimillione	4
African	20	konferenseng	12
Clip	25	Unione	18
Cyril	28	probenseng	46

Although the SPCS is small (contains short code-switched phrases), it does provide a baseline for code-switched corpus for the Sepedi-English language pair. Marivate et al. (2020) created Sepedi news headlines corpus from a national radio news' Facebook page. The data was used to develop a news classification model.

Van Der Westhuizen & Niesler (2016) compiled the spontaneous English-isiZulu code-switched speech corpus from the South African soap operas. The data was manually transcribed and monolingual English text dominated the corpus by 75%. The data was annotated and code-switching boundaries were identified. Multilingual code-switched corpus for English-isiZulu, English-isiXhosa, English-Setswana, and English-Sesotho have been developed (Van Der Westhuizen & Niesler 2018). Data collection for this multilingual code-switched corpus was obtained from digital video recordings of 626 South African soap opera episodes. The data was manually transcribed by the fluent bilingual speakers of the language pairs. The most code-switching trigger words were identified in each of the language pairs.

In another study, Jansen van Vueren & Niesler (2021) used data augmentation to train and evaluate the performance of code-switched language models. Long short-term memory (LSTM) was used as a generative model to synthetically generate code-

switched data to augment the small code-switched datasets. The study observes that optimised models could generate text with an improved perplexity and word-error rate as compared to models without data optimisation that were studied in (Van Der Westhuizen & Niesler 2016).

3 Data collection

The Sepedi radio news and News headlines datasets are used in this study for the development of the models and analisis. The Sepedi radio news dataset is the primary dataset for this study. The data was collected from a community radio station based in Limpopo to create a code-switched corpus. Several broadcast shows are presented daily, and the data collection focused on radio news read during the various times of the day between June and August 2022. Data cleaning was performed to standardise the data. We used the dataset to train the developed Transformer-based text generation model and observed its accuracy when generating synthetic news.

Table 1, shows the size of the datasets with a split of 70% for training, 20% for validation, and 10% for testing". The news headlines dataset Marivate et al. (2020) was used for comparison with the Sepedi radio news dataset. The dataset is relatively small with 1 182 sentences with minimal code-switching.

Table 2, shows the top 10 English and borrowed words in the Sepedi Radio news dataset. The total number of English unique tokens is 136 which constitutes just 8.8% of the total unique tokens in the dataset. Only 9% of English unique tokens have a frequency of 10 and higher. The average number of words per sentence in the combined dataset is 18.2. We randomly sampled 30 sentences in the radio news dataset and observed that English nouns and proper nouns dominated the form of code-switching in the dataset along with the usage of borrowed words as can be seen in Table 2. The number of English words were 31 out of 929 words in the sampled data. This translates to a low frequency of code-switching of just 3.3% with a ratio

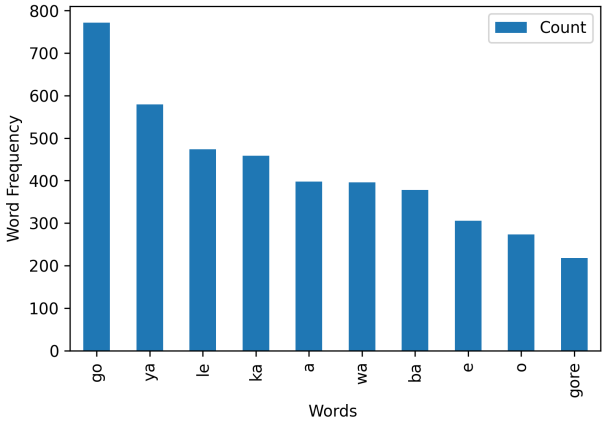

Figure 1: Word frequency in the training dataset

of just 1 English word per sentence. Further analysis of the dataset shows no inter-sentential code-switching.

Data visualisation in all datasets shows that bigrams (character level) have a high frequency. Fig. 1 shows the top 10 word frequencies in the Sepedi radio news dataset. The y-axis depicts the frequency of each word in the dataset while the x-axis shows the words in the vocabulary sorted from most to least frequent.

4 Experimental Setup

For comparison with the previous study by Ramalepe et al. (2022) on monolingual data we adopted the same approach that they used to develop the Transformer based model. The developed model has one Transformer block with causal masking on the attention layers, two separate embedding layers for tokens and a token index with one dense layer with 2 attention heads. We used 64 embedding size for model complexity with the default dropout rate of 0.1. Adam was used as the model optimiser and the rectified linear unit (ReLU) as the activation function. The vocabulary size was set 3k (almost the total number of unique tokens in the combined dataset). Both datasets are combined to observe if there is an improvement in the model's performance as the size of the data increases. The model was trained for 50 epochs due to the small amount of data in the dataset.

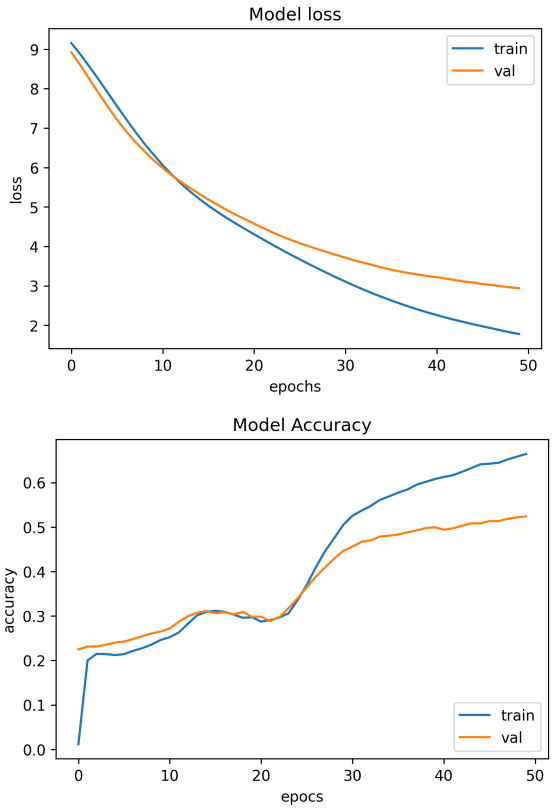

Figure 2: Loss and accuracy curve for the Radio news dataset

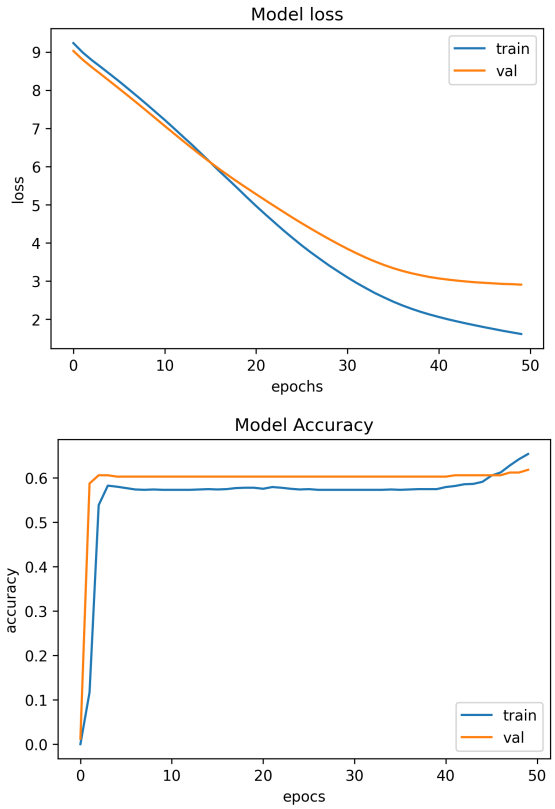

Figure 3: Loss and accuracy curve for the News headlines dataset

The accuracy of the model was measured by computing the total number of correct predictions as a percentage of the total number of predictions during training using the *SparseCategoricalAccuracy* function metric which is used mostly when making text predictions for sparse targets in deep learning. *SparseCategoricalAccuracy* metric checks if the maximum true value is equal to the index of the maximum predicted value.

5 Results and Analysis

Fig. 2 shows the loss and the accuracy curve at each epoch of the training process for the Sepedi radio news dataset, while Fig. 3 shows the loss and the accuracy curve of the news headlines dataset. The loss and accuracy curve for the combined dataset is shown in Fig. 4. The optimal loss rate of the model was 2.361 obtained with the combined dataset with an accuracy of 66%. We observed that although the obtained optimal accuracy was an improvement from 50.3% obtained in Moila & Modipa (2020) using the LSTM based technique on monolingual data, it was still less compared to 75% accuracy obtained in Ramalepe et al. (2022) using the Transformer based approach. Table 3 shows the summary of the results. It is further observed that the size of the dataset influences the accuracy of the model. When the model was trained with the Sepedi radio news dataset(the smallest dataset), the accuracy obtained was low, it increased as we increased the data in the dataset.

In all the figures, Fig. 2, Fig. 3, and Fig. 4 the validation set struggled to generalise, showing an indication of overfitting. This is largely due to the limited amount of data we had in all the datasets. To generate text we start by feeding the model with the starting prompt, the model then generates the conditional probability distribution over the input

Table 3: Validation loss rate and accuracy for each dataset

Name-Dataset	Val-loss error rate	Val-accuracy rate
Radio news	2.938	0.521
News Headlines	3.321	0.61
Combined-datasets	2.361	0.669

Table 4: Generated text

magareng ga tseo di kago letelwa iring ya 13hoo go thobela fm go akaretswa tsa polao ya modiragatsi sibusiso khwinana ba sola anc yo a vandata zinc tša gore magato a gauteng are o mongwe wa marematlou *tsa selegae thekgo hlatlositse* ntwa ya profense ya zululand lehono sa folaga ya matlakadibeseleteng sa go se okobetse ga *pharela ya dienywa tsa citrus go hlola dibaka tsa selegae* go tia ya lehono bodikela mamelodi

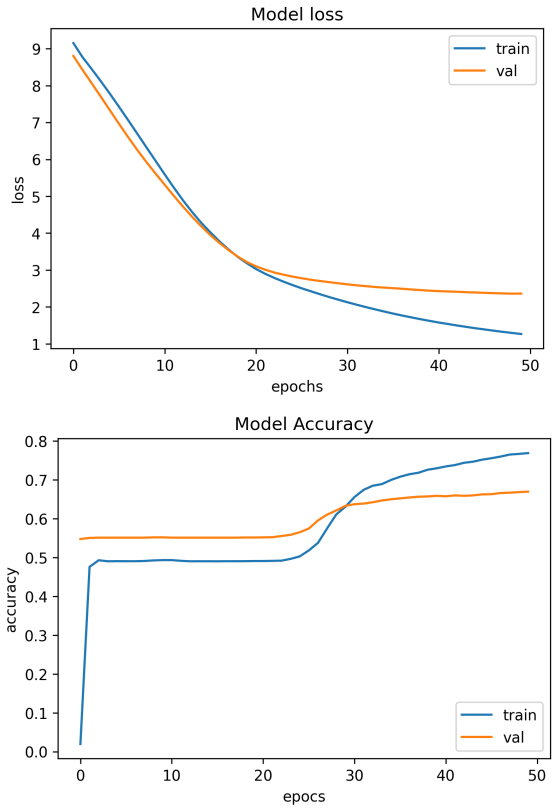

Figure 4: loss and accuracy curve: Combined dataset.

words and sample the next word from the conditional probabilities. The new set of words then becomes the new input to the model, the process continues until the maximum sequence length is reached. To observe the performance of the model on code-switched data we supplied the model with an input text and the model generated the text in Table 4. From the generated text, it can be seen that although the sentences are grammatically correct some are not semantically correct and do not have proper word organisation (formatted in italics in the generated text). For example, the phrase *"tsa selegae thekgo hlatlositse " (of local support increase)* could be corrected as *"hlatlositse thekgo ya tsa selegae"* (increased local support). The number of English words in all the datasets was very small, hence those words did not influence the accuracy of the model and the generated text. However, the performance of the model is low by 12% from the 75% obtained in Ramalepe et al. (2022)

Although the frequency of code-switched words in the datasets was very low, the model could at least generate one code-switched word. For example *"pharela ya dienywa tsa citrus go hlola dibaka tsa selegae"* (The impasse of citrus fruits to create local opportunities). The results signify a positive sentiment of success towards the creation of a large-scale code-switched dataset to train larger models. One major challenge still to be looked at is finding feasible ways of obtaining spontaneous code-switched data as compared to read text. However, such data is often taxing due to the transcription of large amount of speech text.

6 Conclusion

In this study, we discussed the initial phase of collecting, developing, and analysing code-switched corpus using the Sepedi radio news. The developed Sepedi radio news corpus was used to train a Transformer-based text generation model to observe its performance on code-switched data. The model achieved the optimal accuracy of 66% with combined dataset. Data augmentation may be considered in the future to augment the text and create a new code-switched dataset. Other means of collecting spontaneous code-switched data through live recordings and transcription may also be considered in future. To validate the quality of the generated text, human evaluators may also be considered as part of future work.

Acknowledgements

This work is supported by the Centre for Artificial Intelligence Research and the Telkom Centre of Excellence at the University of Limpopo.

References

Chan, J. Y. C., Ching, P. C. & Lee, T. (2005), Development of a Cantonese-English Code-mixing Speech Corpus, *in* 'Ninth European conference on speech communication and technology.'.

Chang, C.-T., Chuang, S.-P. & Lee, H.-Y. (2018), Code-switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation, *in* 'arXiv preprint arXiv:1811.02356.'.

Gao, Y., Feng, J., Liu, Y., Hou, L., Pan, X. & Ma, Y. (2019), Code-switching sentence generation by BERT and generative adversarial networks, *in* 'In INTERSPEECH', Vol. 2019-September, International Speech Communication Association, pp. 3525–3529.

Gupta, D., Ekbal, A. & Bhattacharyya, P. (2020), Findings of the Association for Computational Linguistics A Semi-supervised Approach to Generate the Code-Mixed Text using Pre-trained Encoder and Transfer Learning, *in* 'In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:', pp. 2267–2280.

Hamed, I., Elmahdy, M. & Abdennadher, S. (2017), Building a First Language Model for Code-switch Arabic-English, *in* 'Procedia Computer Science', Vol. 117, Elsevier B.V., pp. 208–216.

Hamed, I., Elmahdy, M., Abdennadher, S., Tagamoa, E. & Khames, E. (2018), Collection and Analysis of Code-switch Egyptian Arabic-English Speech Corpus, *in* 'Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).', pp. 3805–3809.

Jansen van Vueren, J. & Niesler, T. (2021), Optimised Code-Switched Language Model Data Augmentation in Four Under-Resourced South African Languages, *in* 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Vol. 12997 LNAI, Springer Science and Business Media Deutschland GmbH, pp. 303–316.

Lyu, D. C., Tan, T. P., Chng, E. S. & Li, H. (2015), 'Mandarin–English code-switching speech corpus in South-East Asia: SEAME', *Language Resources and Evaluation* **49**(3), 581–600.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R. & Modupe, A. (2020), 'Investigating an approach for low resource language dataset creation, curation and classification: Setswana and Sepedi'.

Modipa, T. I. & Davel, M. H. (2022), 'Two sepedi-english code-switched speech corpora', *Language Resources and Evaluation* **56**(3), 703–727.

Modipa, T. I., Davel, M. H. & de Wet, F. (2013), Implications of Sepedi/English code switching for ASR systems, *in* 'In Proceedings of the Twenty-Fourth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2013)', Johannesburg, South Africa.

Moila, M. M. & Modipa, T. I. (2020), The development of a sepedi text generation model using long-short term memory, *in* 'Proceedings of the 2nd International Conference on Intelligent and Innovative Computing Applications', ACM, New York, NY, USA, pp. 1–5.

Ramalepe, P. S., Modipa, T. I. & Davel, H. M. (2022), The Development of a Sepedi Text Generation Model Using Transformers [Paper presentation], *in* 'SATNAC-2022', George, Western Cape, pp. 51–56.
URL: *https://www.satnac.org.za/proceedings*

Statistics South Africa (2022), Mid-year population estimates 2022, Technical report.
URL: *www.statssa.gov.za,info@statssa.gov.za,Tel+27123108911*

Van Der Westhuizen, E. & Niesler, T. (2016), Automatic Speech Recognition of English-isiZulu Code-switched Speech from South African Soap Operas, Technical report.

Van Der Westhuizen, E. & Niesler, T. (2018), A First South African Corpus of Multilingual Code-switched Soap Opera Speech, *in* 'Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).'.