Low Resource Language Modelling

Ubusuku obuhle namaphupho **amamnandi!**

context

prediction target

Given a sequence of context words, a language model predicts the next word in the sentence. More formally, a language model assigns a probability to a sequence of words. Modern language models are trained on large datasets, however, many of South Africa's languages are low resource – there is little text data available for training language models. We evaluate different language models and training methods for modelling South African languages.

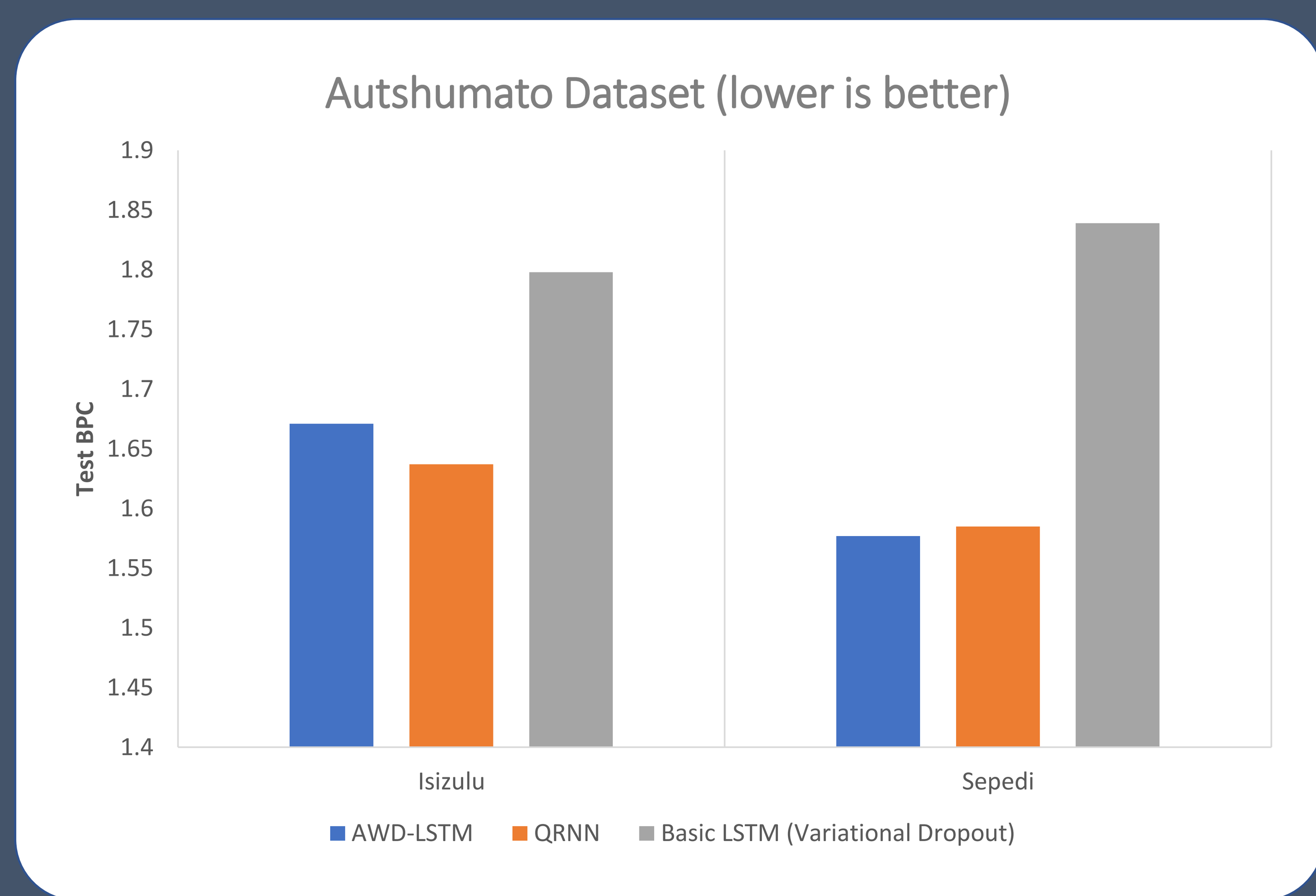
		Hyper-Parameters		
Language	Model	Vocab Size	n-gram order	BPC
isiZulu	n-gram	1000	5	1.832
	FFNNLM	8000	2	**1.815**
Sepedi	n-gram	1000	5	**1.705**
	FFNNLM	10000	2	1.716
	FFNNLM	8000	2	1.716

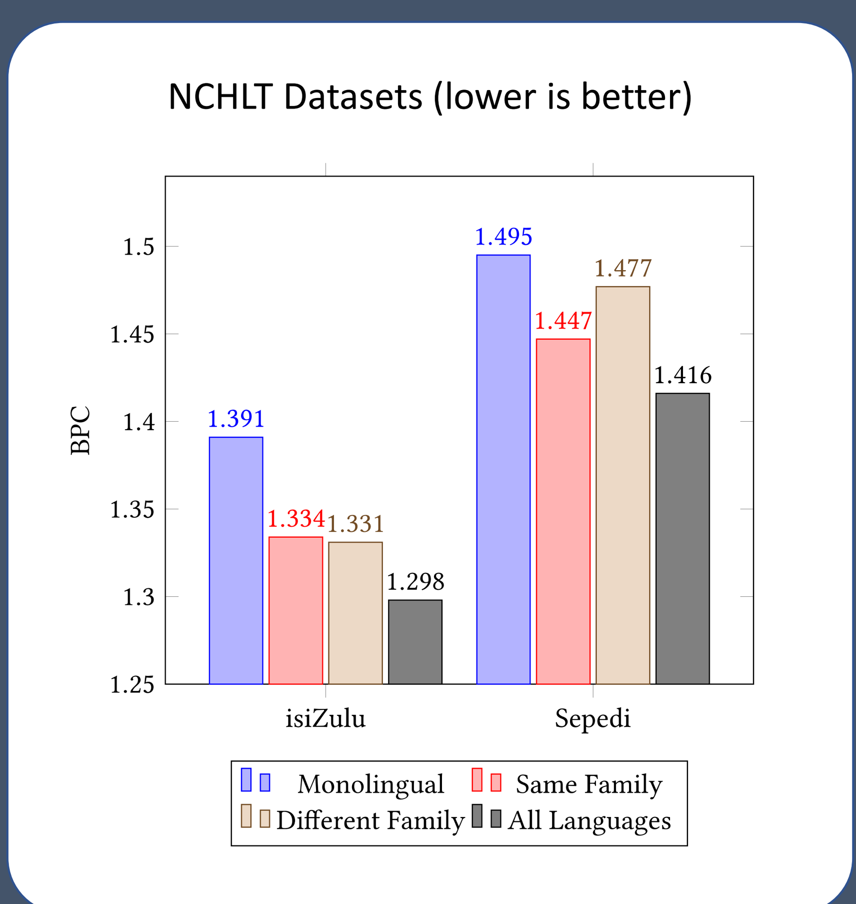
N-Grams

We trained traditional N-gram language models using modified Kneser-Ney smoothing and Feedforward Neural Network language models (FFNNLM) on the low-resource South African Languages isiZulu and Sepedi. We showed that for both the languages of isiZulu and Sepedi, in low-resource conditions, traditional n-gram models and Feedforward Neural Network language models performed similarly in terms of performance. The FFNNLM slightly outperformed the n-gram language model when evaluated on the language isiZulu, however, the n-gram language model slightly outperformed the FFNNLM when evaluated on the language Sepedi.

Recurrent Neural Networks

We tested QRNNs and AWD-LSTMs against a basic LSTM model – on different sized isiZulu and Sepedi Corpora – to determine the effectiveness of the more complex architectures when applied to low resource languages.
Across all datasets the QRNN and AWD-LSTM models notably outperformed the standard LSTM model.





Transformers

We trained Transformer language models on different combinations of low-resource South African languages. We showed that when developing isiZulu and Sepedi language models, training on data from all South African Bantu languages yielded improved performance over training on the individual language or on languages from the same family.

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

Project Team
Stuart Mesham (Transformers)
Jared Shapiro (N-grams)
Luc Hayward (LSTMs)

Supervised by Dr. Jan Buys