_t.png _t.bb _t.png height

Contribution ID: **92** Type: **Student Micro-talk**

Optimised Code-Switched Language Model Data Augmentation in Four Under-Resourced South African Languages

Code-switching in South African languages is common but data for language modelling remains extremely scarce. We present techniques that allow recurrent neural networks (LSTMs) to be better applied as generative models to the task of producing artificial code-switched text that can be used to augment the small training sets. We propose the application of prompting to favour the generation of sentences with intra-sentential language switches, and introduce an extensive LSTM hyperparameter search that specifically optimises the utility of the artificially generated code-switched text. We use these strategies to generate artificial code-switched text for four under-resourced South African languages and evaluate the utility of this additional data for language modelling. We find that the optimised models are able to generate text that leads to consistent perplexity and word error rate improvements for all four language pairs, especially at language switches. We conclude that prompting and targeted hyperparameter optimisation are an effective means of improving language model data augmentation for code-switched speech recognition.

Student?

Yes

Supervisor name

Professor Thomas Niesler

Supervisor email

trn@sun.ac.za

Primary author: JANSEN VAN VUEREN, Joshua (Stellenbosch University)

Co-author: NIESLER, Thomas (University of Stellenbosch)

Presenter: JANSEN VAN VUEREN, Joshua (Stellenbosch University)

Session Classification: Micro-talks

Track Classification: Cognitive Computing and Machine Learning