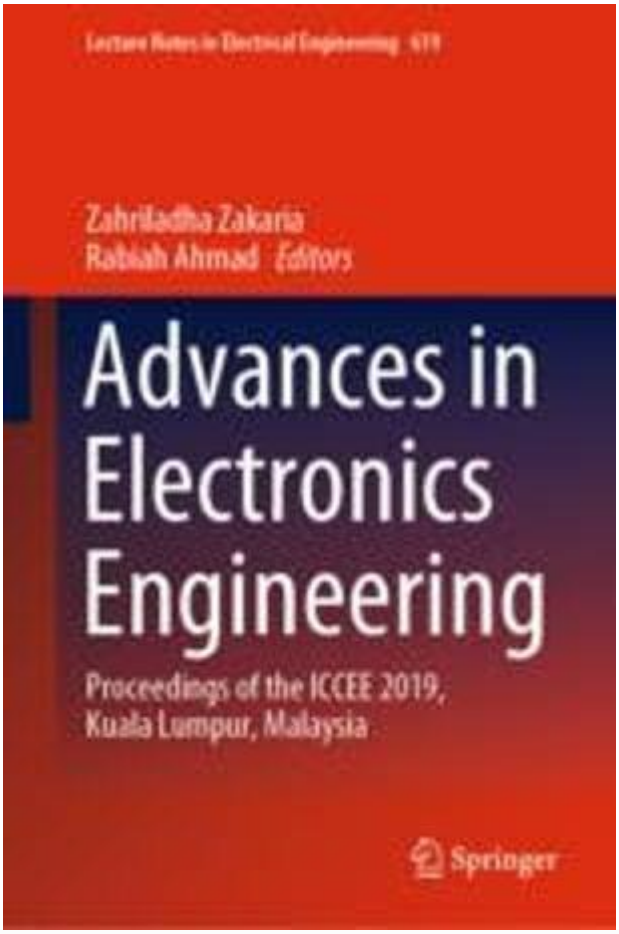

Advances in Electronics Engineering pp 147-158| Cite as

Language Modelling for a Low-Resource Language in Sarawak, Malaysia

- Authors
- Authors and affiliations

- Sarah Samson Juan
- Muhamad Fikri Che Ismail
- Hamimah Ujir
- Irwandi Hipiny

○
1 1.
2 2.

Conference paper
First Online: 17 December 2019

Part of the Lecture Notes in Electrical Engineering book series (LNEE, volume 619)

Abstract

This paper explores state-of-the-art techniques for creating language models in low-resource setting. It is known that building a good statistical language model requires a large amount of data. Therefore, models that are trained on low-resource language suffer from poor performances. We conducted a study on current language modelling techniques such as *n*-gram and recurrent neural network (RNN) to observe their outcomes on data from a language in Sarawak, Malaysia. The target language is Iban, a widely spoken language in this region. We have collected news data form an online source to build an Iban text corpus. After normalising the data, we trained trigram and RNN language models and tested on automatic speech recognition data. Based on our results, we observed that the RNN language models did not significantly outperform the trigram language models. A slight improvement on RNN model is seen after the size of the training data was increased. We have also experimented on merging *n*-gram and RNN language models and we obtained 32.33% improvement after using a trigram-RNN language model.

Keywords

Low-resource language *n*-gram language model Recurrent neural network language model