Evaluation and Compression of Different Language Transformer Models for Semantic Textual Similarity Binary Task Using Minority Language Resources

Authors : Ma. Gracia Corazon Cayanan, Kai Yuen Cheong, Li Sha

Abstract : Training a language model for a minority language has been a challenging task. The lack of available corpora to train and fine-tune state-of-the-art language models is still a challenge in the area of Natural Language Processing (NLP). Moreover, the need for high computational resources and bulk data limit the attainment of this task. In this paper, we presented the following contributions: (1) we introduce and used a translation pair set of Tagalog and English (TL-EN) in pre-training a language model to a minority language resource; (2) we fine-tuned and evaluated top-ranking and pre-trained semantic textual similarity binary task (STSB) models, to both TL-EN and STS dataset pairs. (3) then, we reduced the size of the model to offset the need for high computational resources. Based on our results, the models that were pre-trained to translation pairs and STS pairs can perform well for STSB task. Also, having it reduced to a smaller dimension has no negative effect on the performance but rather has a notable increase on the similarity scores. Moreover, models that were pre-trained to a similar dataset have a tremendous effect on the model's performance scores.

Keywords : semantic matching, semantic textual similarity binary task, low resource minority language,fine-tuning, dimension reduction, transformer models
Conference Title : ICCLNLP 2022 : International Conference on Computational Linguistics and Natural Language Processing
Conference Location : Singapore, Singapore
Conference Dates : May 05-06, 2022

Open Science Index, Computer and Information Engineering Vol:16, No:05, 2022 publications.waset.org/abstracts/145745.pdf