Incremental Learning Static Word Embeddings for Low-Resource NLP

Nathan J. Lee
School of Computer Science, BINUS University, Jakarta, 11480 Indonesia

Nur Afny C. Andryani
Doctor of Computer Science Department, BINUS Graduate Program, BINUS University, Jakarta, 11480 Indonesia

Abstract:

Natural Language Processing (NLP) development for Low-Resource Languages (LRL) remains challenging due to limited data availability, linguistic diversity, and computational constraints. Many NLP solutions rely on complex models and high volume/quality data, which makes them difficult to use in Low-Resource NLP. Inspired by the NLP challenges and insights revealed by various previous works, the underexplored Incremental Learning (IL) Static Word Embedding (SWE) system to the test in the low-resource NLP case of Indonesia's local languages is proposed and presented. With basic-level models and hyperparameter sweeps, these models are tested in the scenario of incrementally incorporating 10 different local languages into themselves. The simulations indicate this type of model resists Catastrophic Forgetting (CF) very well and delivers competitive performance on the downstream task of sentiment analysis. In terms of f1 scores, the proposed model succeeds to exceed other baseline models and even rival heavy Transformer models. The proposed model can be considered as a prospective holistic solution for low-resource NLP. Future works could explore this model's behavior in finer-grained NLP tasks, different IL settings, or test more advanced models.

Keywords:

Incremental Learning, Indonesian, Low Resource, NLP, Sentiment Analysis, Static Word Embedding.