Bridging Linguistic Divides: Cross-Lingual Transfer Learning and Pseudo-Labeling for Natural Language Processing in Low-Resource Languages

Granucci, Lucas (School: Minnetonka High School)

The lack of research in artificial intelligence systems for parsing and understanding text in less commonly spoken languages has led to a barrier to access to state-of-the-art natural language processing (NLP) models. This experiment uses cross-lingual transfer learning, as well as self-training via pseudo labeling, to improve the performance of named entity recognition (NER) models in low-resource languages, demonstrating that these novel techniques are effective at reducing the problems faced by limited data accessibility in underrepresented languages. To most closely replicate the results of cutting-edge NER models, different deep-learning model architectures were developed, trained, and evaluated on languages from six different language families. BERT-BiLSTM-CRF achieved the highest average F1-score of 87%, outperforming the BERT, BERT-BiLTM, and BERT-CRF architectures by 4.315%, 4.766%, and 0.627% respectively. Cross-lingual transfer learning was also tested to see if the knowledge a model gained from a syntactically similar language would transfer over to languages with less data available. Results show that model performance improves by an average of 5.49% across all language groups, with 13.00% being the largest improvement. Pseudo-labeling was also tested and shown to improve performance by as much as 6.99% without requiring additional manually labeled data. This study has shown that these methods are effective at improving performance of NLP models for low-resource languages, demonstrating their potential to bridge the language gap between high and low resource languages.

Awards Won: