Genealogical Approach to Low-Resource Language Modelling

Brandon Chaperon, Suheng Qian

Department of Linguistics, Research Mentor: Siva Reddy

Introduction

- mBERT underperforms on low-resource languages
- Historical linguists believe some languages are related to each other
- We experiment a genealogical approach to improving a low-resource language model

Key findings

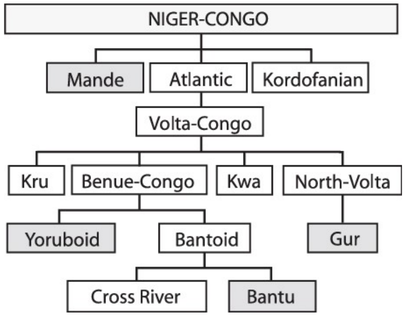
- mBERT is able to improve its performance on a language when trained on a dataset coming from another related language
- A shortcut for researchers without much data for a specific language

Methods

- Perform certain tasks on Wolof with mBERT that was fine-tuned individually on Yoruba, Swahili, and Amharic in order to compare with the original mBERT model
- Name-Entity Recognition (NER)
- Part-of-speech (POS) tagging

Data Analysis

- mBERT fine-tuned on Yoruba, Swahili, and Amharic improved those models by ~3% on NER



- We hypothesized that mBERT trained on a closely related language would improve the performance on Wolof. We also thought that an unrelated language would worsen the model.

Results

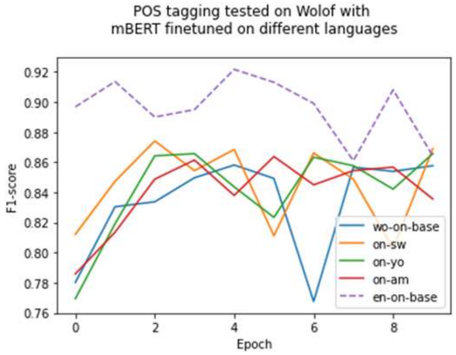
Performance after 10 epochs

Finetuning	POS	NER
Baseline	85.77	93.43
Swahili	86.90	91.29
Amharic	83.56	91.78
Yoruba	86.57	91.17

Mean performance of 10 epochs

Finetuning	POS	NER
Baseline	83.37	93.31
Swahili	84.54	94.93
Amharic	84.03	93.30
Yoruba	84.14	88.78

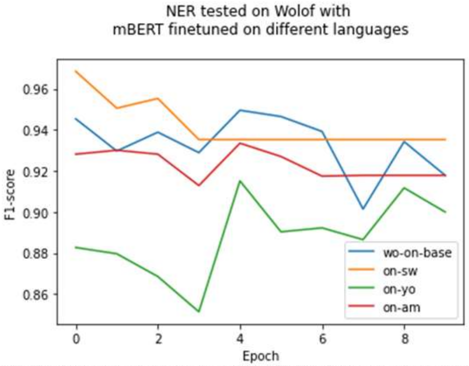
- On a baseline mBERT model, English does 89% on POS tagging



- Using these models, we were able to improve the performance of POS tagging on Wolof using Yoruba and Swahili and worsen it with Amharic
- On the other hand, there was no difference on NER tasks

Conclusions

- Fine-tuning mBERT on related languages is useful to improve the performance of mBERT on a related language
- This can be used for low-resource languages
- But it appears to only work on syntactic-related tasks.



Datasets

- MasakhaNER by Masakhane
- POS tagging by Universal Dependencies

McGill UNIVERSITY