IgboBERT Models: Building and Training Transformer Models for the Igbo Language

Chiamaka Chukwuneke, Ignatius Ezeani, Paul Rayson and Mahmoud El-Haj

Abstract

This work presents a standard Igbo named entity recognition (IgboNER) dataset as well as the results from training and fine-tuning state-of-the-art transformer IgboNER models. We discuss the process of our dataset creation - data collection and annotation and quality checking. We also present experimental processes involved in building an IgboBERT language model from scratch as well as fine-tuning it along with other non-Igbo pre-trained models for the downstream IgboNER task. Our results show that, although the IgboNER task benefited hugely from fine-tuning large transformer model, fine-tuning a transformer model built from scratch with comparatively little Igbo text data seems to yield quite decent results for the IgboNER task..

Keywords: Igbo, named entity recognition, BERT models, under-resourced, dataset.

Introduction

- Igbo is a south-eastern Nigerian language and belongs to Niger-Congo family.
- One of 3 major official Nigerian languages and official minority language in Equatorial Guinea and Cameroon.
- Approximately 45 million speakers but have few NLP resources(low-resourced) for interaction in this technological era.
  - Created human-labelled publicly-available datasets for Igbo language
  - Built an IgboBERT language model from scratch.
  - Released the code, data and model publicly to support Igbo NLP and also to contribute to the African NLP efforts.

Challenges of Igbo language

- Orthography: Igbo text corpora are written with combination of Lepsuis, Africa and Ọnwụ orthographies.
- Ambiguity: Some Igbo words are relatively ambiguous. For instance, some person names have other meanings, e.g. "`Eze`" can be the name of a person (proper noun), a part of the human body for chewing (plural noun) and also it can be a male ruler of an independent state (noun).

Data Collection

Source	Sentences)	Tokens	Orthography
eze-goes-to-school.txt	1272	25413	Ọnwụ
mmadu-ka-a-na-aria.txt	2023	39731	Ọnwụ
bbc-igbo.txt	34056	566804	Africa, Ọnwụ
igbo-radio.txt	5131	191450	Lepsuis, Africa, Ọnwụ
jw-ot-igbo.txt	32251	712349	Lepsuis, Ọnwụ
jw-nt-igbo.txt	10334	253806	Lepsuis, Ọnwụ
jw-books.txt	142753	1879755	Lepsuis, Ọnwụ
jw-teta.txt	14097	196818	Lepsuis, Ọnwụ
jw-ulo-nche.txt	27760	392412	Lepsuis, Ọnwụ
jw-ulo-nche-naamu.txt	113772	1465663	Lepsuis, Ọnwụ
igbo-radio.txt	2120	11173	Lepsuis, Africa, Ọnwụ
kaoditaa.txt	5880	22557	Lepsuis, Africa, Ọnwụ
Total	391,449	5757931	

Table 1: Data Sources and counts

NER Annotation

- Used BIO tagging scheme to label the entities..
- Personal name (PER), organization (ORG), location (LOC), date and time (DATE) was tagged.
- The annotation of the IgboNER was performed using the ELISA tool.

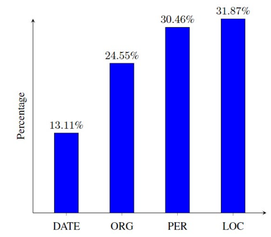

Figure 1: Entity distribution. This shows the percentage distribution of the entities annotated.

Experiment and Result

- An Igbo language model (IgboBERT) was trained from scratch using transformers and tokenizers to have a baseline model for Igbo language NER.
- IgboBERT was pre-trained at a learning-rate of 1e-4 for 5 epoch with a masked language modeling (MLM) objective.
- We then fine-tuned the IgboBERT model on NER downstream task using our IgboNER dataset.
- mBERT, XLM-RoBERTa, DistillBERT were also fine-tuned to a downstream NER task using the IgboNER dataset.

Model	Precision	Recall	F1	Accuracy
mBERT	85.67	87.67	86.66	97.96
XLM-R	84.54	85.67	85.10	97.81
DistilBERT	79.79	77.00	78.37	96.20
IgboBERT	76.44	79.50	77.94	95.61

Figure 3: Displays the fine-tuned results of mBERT, XML-R, DistilBERT and IgboBERT models after 20 epoch at 1e-4 learning rate.

Conclusion and Summary

- The first IgboNER dataset
- The first and only transformer-based language model pre-trained on the Igbo language was developed.
- IgboBERT achieved an F1 of 77.94 but there was no convergence in the training vs. validation loss (over-fitting).
- The issue of over-fitting among other research directions will be handled in our further studies by automatically creating more IgboNER datasets and also use gazetteers to increase our dataset.

Figure 2: Link.

Acknowledgement

We acknowledge the support of Tertiary Education Trust Fund, TETFund,Nigeria for funding this research.We also acknowledge JW.org, Igbo-Radio and Kaoditaa who gave us permission to use data from their websites. We will not forget to thank the anonymous reviewers for their useful comments and for reading the draft.