Investigating the quality of pre-trained word embeddings in Yorùbá, a low-resource Language

Jesujoba Oluwadara Alabi[1,3] David Ifeoluwa Adelani[2,3]

jesujoba_oluwadara.alabi@dfki.de

didelani@lsv.uni-saarland.de

[1]German Research Center for Artificial Intelligence (DFKI)

[2]Spoken Language Systems (LSV), Saarland Informatics Campus,

[3]Saarland University, Saarbrücken, Germany

Aim

In this paper, we investigate the quality of one of the most widely used multilingual word embedding, FastText.

We show that the word embeddings do not carry so much semantic and syntactic relationship as expected. We found out that the text data used for training are of low quality.

Yorùbá Language

- Yorùbá is a language in the West Africa with over 50 million speakers. It is spoken among other languages in Nigeria, republic of Togo, Benin Republic, Ghana and Sierra Leone.

- The Standard Yorùbá has 25 alphabets without the Latin letters c, q, v, x and z. There are three tones represented with tonal symbols in Yorùbá namely low(\), mid(−) and high(/). These tones are applied on vowels and syllabic nasals.

Experimental Setup

- We compare the cosine similarities of two pre-trained FastText embeddings (trained on Yorùbá Wiki and Common Crawls & Wiki) and FastText embeddings trained on (high-quality) Curated data.

- For evaluation, we **translated WordSim-353 dataset** from English to Yorùbá and compare the performance of the embeddings to the human similarity scores (with scores from 0 to 10).

- To measure the **quality of embeddings**, we compute the *Spearmanr Correlation* between human similarity scores and the cosine similarities of the Yorùbá word pairs.

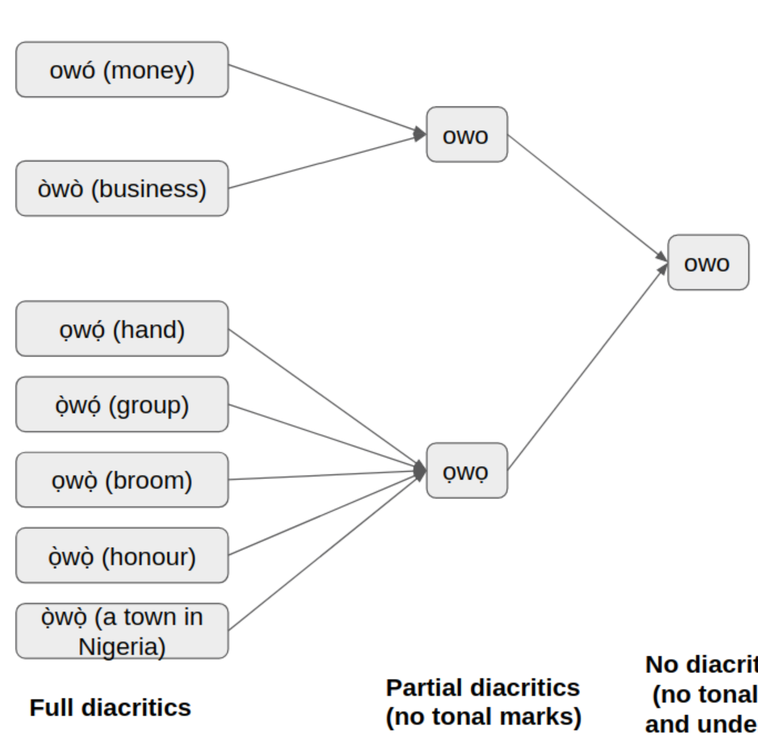
Dataset

We curated a small dataset of 1.2 million tokens that are of high-quality (or clean) Yorùbá text. Also, we compare the impact of adding more clean texts (especially from JW300) and noisy texts (i.e texts without/inconsistent tonal symbols).

Description	Source URL	#tokens	status
Lagos-NWU	rma.nwu.ac.za	24,868	clean
Yorùbá Bible	www.bible.com	819,101	clean
GlobalVoices	yo.globalvoices.org	24,617	clean
JW news	www.jw.org/yo	170,203	clean
Ìrìnkèrindò	manual	56,434	clean
Igbó Olódùmarè	manual	62,125	clean
JW300 Yorùbá	opus.nlpl.eu	10,558,055	clean
Yorùbá Tweets	Niger-Volta-LTI Github	153,716	clean
BBC Yorùbá	bbc.com/yoruba	330,490	noisy
VON Yorùbá	von.gov.ng/yoruba/	380,252	noisy
Yorùbá Wiki Subset	yo.wikipedia.org	129,075	noisy

Summary figures of the corpora used in the analysis.

Yorùbá Diacritics Problem



Most texts on the web are not written with proper diacritics

Results

Model	Vocab Size	#Pairs Found	Spearmanr
Pre-trained Model (Wiki)	21,730	350	0.136
Pre-trained Model (Common Crawl & Wiki)	151,125	242	0.073
Curated *Small* Dataset (Clean, Full diacritics)	12,268	353	0.322
Curated *Small* Dataset (Partial diacritics)	10,593	353	0.316
Curated *Large* Dataset (Clean, Full diacritics)	31,339	353	0.387
Curated *Large* Dataset (Partial diacritics)	27,558	353	0.377
Curated *Large* Dataset (Clean + Noisy)	44,560	353	**0.391**

Spearmanr Correlation between human evaluation and embedding similarities. For the pre-trained Model trained on Common crawls and wikipedia, we found out that over 135,000 tokens are not Yorùbá tokens using a *langdetect* python tool. This shows that the quality of data used for training is very poor.

Examples

Eng1	Eng2	Yor1	Yor2	Human	Pre-trained Wiki	Pre-trained CC & Wiki	Curated Large Full Diacritics
professor	student	ọ̀jọ̀gbọ́n	akẹ́ẹ̀kọ́	6.81	-0.13	NA	1.91
law	lawyer	òfin	agbẹjẹ́rò	8.38	5.28	-0.07	2.01
king	queen	ọba	ayaba	8.58	6.83	0.17	4.04
Jerusalem	Israel	jerúsálẹ́mù	ísráẹ́lì	8.46	2.16	NA	3.12
drink	eat	mu	jẹ	6.87	-3.36	-0.04	2.0
money	wealth	owó	ọrọ̀	8.27	0.86	NA	2.88

Sample of word similarity scores(from 0 to 10) obtained from the trained models using WS-353 dataset