

Yorùbá and beyond: NLP from an African perspective

Jesujoba O Alabi

Outline

- Introduction
- Yorùbá
- Our recent work
- BERT for Low Resource Languages
- Other Initiatives

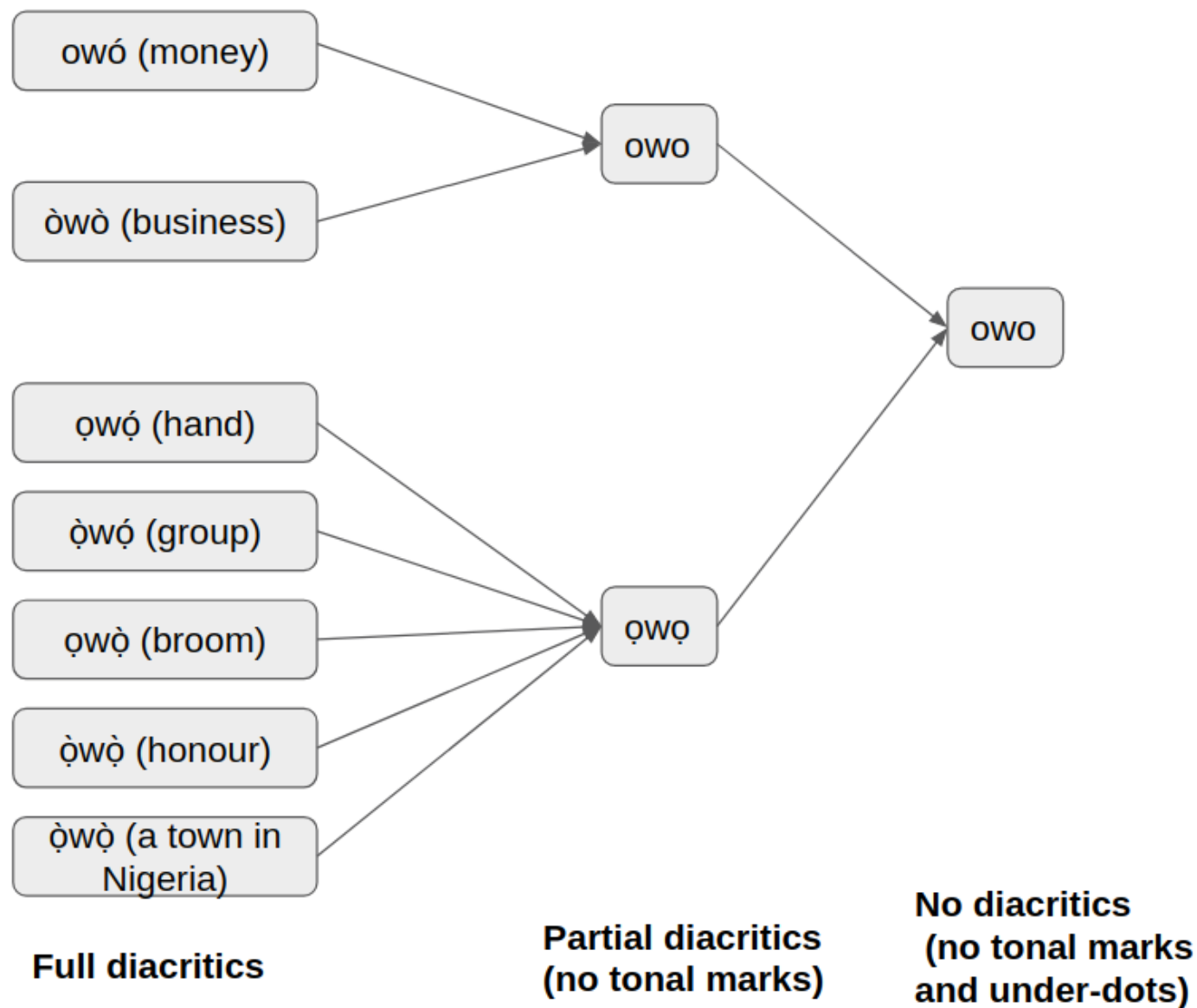
Introduction

- Africa has a lot of languages
- 1500-2000 languages
- 1/3 of world languages
- typically low resource languages
- Yorùbá is the 3rd most spoken language.
- Niger Congo – (benue-congo)
- Uses diacritics



Yorùbá

- Most of the Yorùbá texts found online either use
 - correct Yorùbá orthography or
 - replace diacritized characters Ashiah (2014)
- Yorùbá text in the public domain today is not well diacritized.
- Wikipedia is not an exception.
- No evaluation dataset
- Small Corpora



Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorùbá and Twi

Jesujoba O. Alabi^{*†‡} Kwabena Amponsah-Kaakyire^{*†‡} David I. Adelani^{†‡} Cristina España-Bonet^{†‡}

[†]DFKI GmbH, Saarbrücken, Germany

[‡]Spoken Language Systems (LSV), Saarland Informatics Campus, [‡]Saarland University, Saarbrücken, Germany

{jesujoba.oluwadara.alabi, kwabena.amponsah-kaakyire, cristinae}@dfki.de, didelani@lsv.uni-saarland.de

Abstract

The success of several architectures to learn semantic representations from unannotated text and the availability of these kind of texts in online multilingual resources such as Wikipedia has facilitated the massive and automatic creation of resources for multiple languages. The evaluation of such resources is usually done for the high-resourced languages, where one has a smorgasbord of tasks and test sets to evaluate on. For low-resourced languages, the evaluation is more difficult and normally ignored, with the hope that the impressive capability of deep learning architectures to learn (multilingual) representations in the high-resourced setting holds in the low-resourced setting too. In this paper we focus on two African languages, Yorùbá and Twi, and compare the word embeddings obtained in this way, with word embeddings obtained from curated corpora and a language-dependent processing. We analyse the noise in the publicly available corpora, collect high quality and noisy data for the two languages and quantify the improvements that depend not only on the amount of data but on the quality too. We also use different architectures that learn word representations both from surface forms and characters to further exploit all the available information which showed to be important for these languages. For the evaluation, we manually translate the wordsim-353 word pairs dataset from English into Yorùbá and Twi. We extend the analysis to contextual word embeddings and evaluate multilingual BERT on a named entity recognition task. For this, we annotate with named entities the Global Voices corpus for Yorùbá. As output of the work, we provide corpora, embeddings and the test suits for both languages.

Keywords: Multilingual embeddings, Low-resource language, Yorùbá, and Twi

Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorùbá and Twi

01

Investigate **quality** of word embeddings on two African languages:

- FastText & BERT

02

Compare pre-trained embeddings & our trained embeddings

03

Analyze the impact of adding noisy-texts (low-quality) to high quality curated dataset.

04

Learn representations using word and sub-word representations:

- FastText & CWE

Description	Source URL	#tokens	Status	C1	C2	C3
<i>Yorùbá</i>						
Lagos-NWU corpus	github.com/Niger-Volta-LTI	24,868	clean	✓	✓	✓
Alákòwé	alakoweyoruba.wordpress.com	24,092	clean	✓	✓	✓
Òrò Yorùbá	oroyoruba.blogspot.com	16,232	clean	✓	✓	✓
Èdè Yorùbá Rẹwà	deskgram.cc/edeyorubarewa	4,464	clean	✓	✓	✓
Doctrine \$ Covenants	github.com/Niger-Volta-LTI	20,447	clean	✓	✓	✓
Yorùbá Bible	www.bible.com	819,101	clean	✓	✓	✓
GlobalVoices	yo.globalvoices.org	24,617	clean	✓	✓	✓
Jehova Witness	www.jw.org/yo	170,203	clean	✓	✓	✓
Ìrìnkèrindò nínú igbó elégbèje	manual	56,434	clean	✓	✓	✓
Igbó Olódùmarè	manual	62,125	clean	✓	✓	✓
JW300 Yorùbá corpus	opus.nlpl.eu/JW300.php	10,558,055	clean	✗	✗	✓
Yorùbá Tweets	twitter.com/yobamoodua	153,716	clean	✓	✓	✓
BBC Yorùbá	bbc.com/yoruba	330,490	noisy	✗	✓	✓
Voice of Nigeria Yorùbá news	von.gov.ng/yoruba	380,252	noisy	✗	✗	✓
Yorùbá Wikipedia	dumps.wikimedia.org/yowiki	129,075	noisy	✗	✗	✓
<i>Twi</i>						
Bible	www.bible.com	661,229	clean	✓	✓	✓
Jehovah's Witness	www.jw.org/tw	1,847,875	noisy	✗	✗	✓
Wikipedia	dumps.wikimedia.org/twwiki	5,820	noisy	✗	✓	✓
JW300 Twi corpus	opus.nlpl.eu/JW300.php	13,630,514	noisy	✗	✗	✓

Table 1: Summary of the corpora used in the analysis. The last 3 columns indicate in which dataset (C1, C2 or C3)

i. Curated Small Dataset (clean), C1

- i. Yorùbá: 1.6 million tokens
- ii. Twi: 735k tokens

ii. Curated Small Dataset (clean + noisy), C2 (Wikipedia, BBC Yorùbá)

- i. Yorùbá: 2 million tokens
- ii. Twi: 742k tokens

iii. Curated Large Dataset, C3

Word Embeddings

- Word embeddings have been proven to be very useful for training downstream natural language processing (NLP) tasks.
- Contextualized have been shown to further improve the performance of NLP.

Task	Dataset	Model	Metric
Word Similarity	Translated WordSim-353	FastText, CWE	Spearman Correlation
Named Entity Recognition	Global Voices News Yorùbá Dataset	BERT	F1-Score

Evaluation on FastText

Model	Twi		Yorùbá	
	Vocab Size	Spearman ρ	Vocab Size	Spearman ρ
F1: Pre-trained Model (Wiki)	935	0.143	21,730	0.136
F2: Pre-trained Model (Common Crawl & Wiki)	NA	NA	151,125	0.073
C1: Curated <i>Small</i> Dataset (Clean text)	9,923	0.354	12,268	0.322
C2: Curated <i>Small</i> Dataset (Clean + some noisy text)	18,494	0.388	17,492	0.302
C3: Curated <i>Large</i> Dataset (All Clean + Noisy texts)	47,134	0.386	44,560	0.391

Table 2: FastText embeddings: Spearman ρ correlation between human judgements and similarity scores on the wordSim-353 for the three datasets analysed (C1, C2 and C3)

BERT Evaluation on NER Task

Entity type	Number of tokens			
	Total	Train	Val.	Test
ORG	289	214	40	35
LOC	613	467	47	99
DATE	662	452	86	124
PER	688	469	109	110
O	23,988	17,819	2,413	4,867

Table 3: Number of tokens per named entity type in the Global Voices Yorùbá corpus.

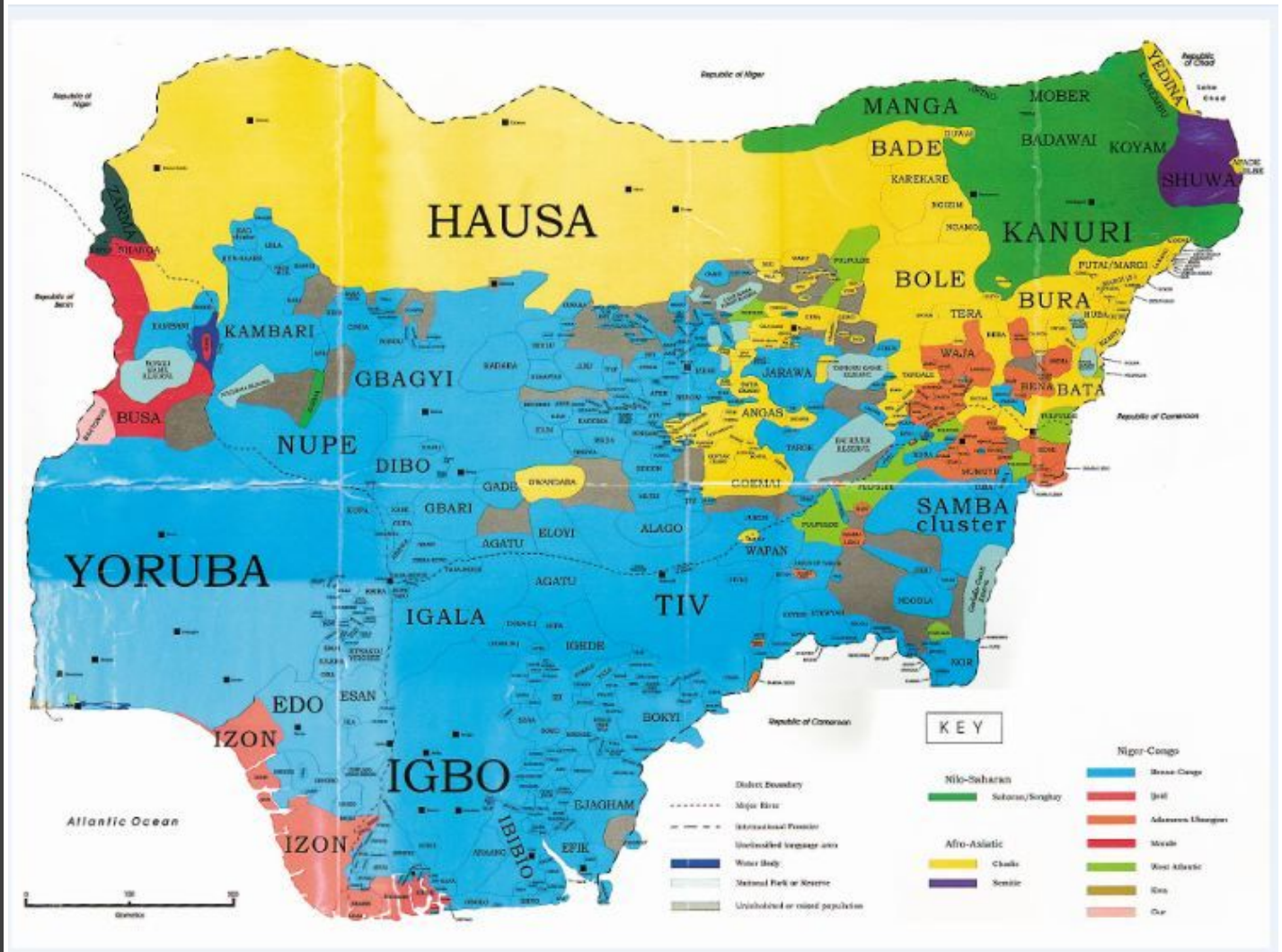
Embedding Type	DATE	LOC	ORG	PER	F1-score
Pre-trained <i>uncased</i> Multilingual-bert (Multilingual vocab)	44.6	33.9	12.1	5.7	27.1 \pm 0.7
Fine-tuned <i>uncased</i> Multilingual-bert (Multilingual vocab)	64.0	65.3	38.8	47.4	56.4 \pm 2.4
Fine-tuned <i>uncased</i> Multilingual-bert (Yorùbá vocab)	67.0	71.5	40.4	49.4	60.1 \pm 0.8

Table 4: NER F1 score on Global Voices Yorùbá corpus after fine-tuning BERT for 10 epochs. Mean F1-score computed after 5 runs

BERT and other LMs

- Lauscher et al.(2020) find that the transfer for multilingual trans-former models is less effective for resource-lean settings and distant languages.
- Fast adaptation method for obtaining a bilingual BERT of English and a target language. Tran (2020)
- Where target language could be any African Language.
- Just mono lingual data is needed.

- Over 500 spoken languages
- Yoruba, Hausa and Igbo are major
- Can we have BERT for all possible Nigerian Languages?
- Can we have BERT for Nigerian Languages from the same class?
- BERT in low resource setting.



Way to go

- Replicate our work for other African languages
- For example, Xhosa
 - Xhosa has a lot of resources online (JW300, OPUS, Common crawl, etc)
 - Limited Wikipedia articles
 - No diacritics
 - No word embeddings!

Other ways to go

- Standardization of existing dataset – e.g. BBC Yoruba, Wikipedia, VON
- Automatic Diacritics Application – NN Based
- Wikipedia articles writing/translation

Thank you for Listening!

References

- Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavas. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. ArXiv, abs/2005.00633.
- Ke Tran, 2020. From English to Foreign Languages: Transferring Pretrained Language Models. ArXiv, abs/2002.07306..
- Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina Espana-Bonet. 2020. Massivevs. Curated Word Embeddings for Low-Resourced Languages. The Case of Yor`ub`a and Twi. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 2747–2755, Marseille, France. European Language Resources Association.
- <https://translatorswithoutborders.org/language-data-nigeria/>
- https://upload.wikimedia.org/wikipedia/commons/thumb/c/c4/African_language_families_en.svg/553px-African_language_families_en.svg.png
- <https://i.pinimg.com/originals/00/13/f4/0013f4b94fa1b2b85221f9e22fd43b8c.jpg>
- https://www.nationsonline.org/oneworld/african_languages.htm