Improving Yorùbá Diacritic Restoration

Iroro Fred Ọ̀nọ̀mẹ̀ Orife[1] David I. Adélaní[1,4] Timi Fasubaa[1]
Victor Williamson[5] Wuraola Fisayo Oyewusi[3] Ọlámilékan Wahab[1] Kọ́lá Túbọ̀sún[2]

[1]Niger-Volta Language Technologies Institute,
[2]Yorùbá Name, [3]Data Science Nigeria,
[4]Saarland University, [5]University of Wisconsin-Milwaukee

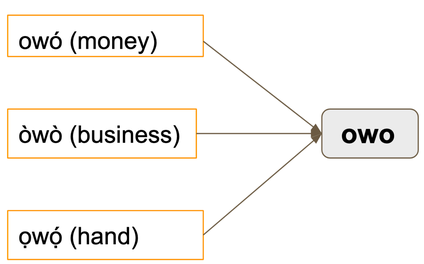
The Yorùbá Language

- 40M speakers in Nigeria & West Africa
- A **tonal** language
- Uses the Latin alphabet with diacritics
- Non-diacritized text is **useless** for computing

Yorùbá text must be correctly represented in computing environments, to build a robust ecosystem of Yorùbá-firstlanguage technologies including NLP, ASR & TTS applications.

Ambiguous non-diacritized text

Problem:



Automatic diacritic restoration (ADR) systems facilitate text entry and correction that promotes the correct orthography and quotidian usage of the language in electronic media.

Paper (arXiv)
2003.10564

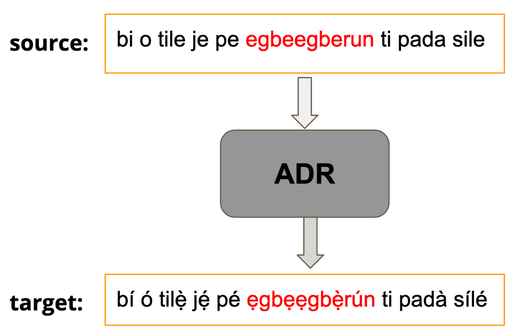
Datasets

Data sources, prevalence and category of text

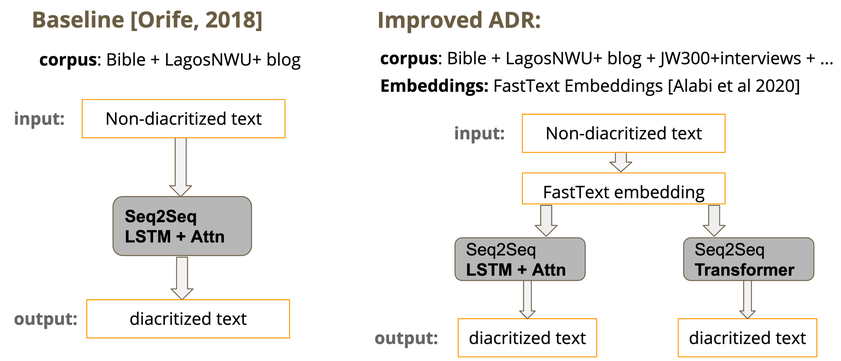
# words	Source	Description
24,868	rma.nwu.ac.za	Lagos-NWU corpus
50,202	theyorubablog.com	language blog
910,401	bible.com/versions/911	Biblica (NIV)
11,488,825	opus.nlpl.eu	JW300
831,820	bible.com/versions/207	KJV
177,675	GitHub	Embeddings dataset
142,991	GitHub	Language ID corpus
47,195		Yorùbá lexicon
29,338	yoruba.unl.edu	Proverbs
2,887	unicode.org/udhr	Human rights edict
150,360	Private sources	Conversations
15,281	Private sources	Short stories
20,038	OCR	Háà Ènìyàn (Fiction)
28,308	yo.globalvoices.org	Global Voices news

Approach

Training idea: Predict diacritics based on context



This Work



Results

BLEU scores, predicted perplexity & WER on the Global Voices testset

Model	BLEU	Perplexity	WER%
Soft-attention model from (Orife, 2018)	26.53	1.34	58.17
+ Language ID corpus	42.52	1.69	33.03
++ Interview text	42.23	1.59	32.58
+ All new text *minus JW300*	43.39	1.60	31.87
+ All new text	**59.55**	1.44	**20.40**
+ All new text with FastText embedding	58.87	**1.39**	21.33
Transformer model			
+ All new text *minus JW300*	45.68	1.95	34.40
+ All new text	59.05	**1.40**	23.10
+ All new text + FastText embedding	**59.80**	1.43	**22.42**

Code (Github)
Niger-Volta-LTI/yoruba-adr