Attentive Sequence-to-Sequence Learning for Diacritic Restoration of Yorùbá Language Text

Iroro Fred Ọ̀nọ̀mẹ̀ Orife
Niger-Volta Language Technologies Institute

Contact Information:
github.com/niger-volta-LTI
iroro@alumni.cmu.edu

Abstract

Yorùbá is a widely spoken West African language with a writing system rich in tonal and orthographic diacritics. With very few exceptions, diacritics are omitted from electronic texts, due to limited device and application support. Diacritics provide morphological information, are crucial for lexical disambiguation, pronunciation and are vital for any Yorùbá text-to-speech (TTS), automatic speech recognition (ASR) and natural language processing (NLP) tasks. Reframing Automatic Diacritic Restoration (ADR) as a machine translation task, we experiment with two different attentive Sequence-to-Sequence neural models to process undiacritized text. On our evaluation dataset, this approach produces diacritization error rates of less than 5%. We have released pre-trained models, datasets and source-code as an open-source project to advance efforts on Yorùbá language technology.

Introduction

Yorùbá is a tonal language spoken by more than 40 Million people in the countries of Nigeria, Benin and Togo in West Africa. There are an additional million speakers in the African diaspora, making it the most broadly spoken African language outside Africa.

On modern computing platforms, the vast majority of Yorùbá text is written in plain ASCII, without diacritics. This presents grave problems for usage of the standard orthography via electronic media, which has implications for the unambiguous pronunciation of Yorùbá's lexical and grammatical tones by both human speakers and TTS systems. Improper handling of diacritics also degrades the performance of document retrieval via search engines and frustrates every kind of Natural Language Processing (NLP) task, notably machine translation to and from Yorùbá. Finally, correct diacritics are mandatory in reference transcripts for any Automatic Speech Recognition (ASR) task.

1. We propose two different NMT approaches, using soft-attention and self-attention sequence-to-sequence (seq2seq) models [1, 2], to rectify undiacritized Yorùbá text.
2. Datasets, pre-trained models and source code are an open-source project at **github.com/niger-volta-LTI/yoruba-adr**

Ambiguity in Undiacritized Yorùbá text

Automatic Diacritic Restoration (ADR), which goes by other names such as Unicodification or deASCIIfication is a process which attempts to resolve the ambiguity present in undiacritized text. Undiacritized Yorùbá text has a high degree of ambiguity. Adegbola et al. state that for ADR the "prevailing error factor is the number of valid alternative arrangements of the diacritical marks that can be applied to the vowels and syllabic nasals within the words".

For our training corpus of 1M words, we quantify the ambiguity by the percentage of all words that have diacritics, 85%; the percentage of unique non-diacritized word types that have two or more diacritized forms, 32%, and the lexical diffusion or *LexDif* metric, which conveys the average number of alternatives for each non-diacritized word, 1.47.

Characters		Examples
à á ǎ	**a**	gbà *(spread)*, gba *(accept)*, gbá *(hit)*
è é ẹ ẹ̀ ẹ́	**e**	esé *(cat)*, èsè *(dye)*, ẹsẹ̀ *(foot)*
ì í	**i**	ìlú *(town)*, ilu *(opener)*, ìlù *(drum)*
ò ó ọ ọ̀ ọ́ ǒ	**o**	ọkó *(hoe)*, ọ̀kọ̀ *(spear)*, ọkọ̀ *(vehicle)*
ù ú ǔ	**u**	mu *(drink)*, mù *(sink)*, mú *(sharp)*
ǹ ń n̄	**n**	n *(I)*, ń (continuous aspect marker)
ṣ	**s**	sá *(run)*, ṣá *(fade)*, ṣà *(choose)*

Table 1: Diacritized forms for each non-diacritic character

Further, 64% of all unique, non-diacritized monosyllabic words possess multiple diacritized forms. When we consider the distribution of ambiguity over grammatical function, we recognize the added difficulty of tasks like the lexical disambiguation of non-diacritized Yorùbá verbs, which are predominantly monosyllabic.
Finally, there are tonal changes, which are rules about how tonal diacritics on a specific word are *altered* based on context.

Verb	Phrase	Translation	Tone
tà	o tà a	he sells it	Low
	o ta iṣu, o taṣu	he sells yams	Mid
tà	o ta ìwé	he sells books	Mid
	o tàwé	he sells books	Low

Table 2: Tonal Changes

A Sequence-to-Sequence Approach

Expressing ADR as a machine translation problem, we treat undiacritized text and diacritized text as source and target languages respectively in a NMT formulation. We experimented two different NMT approaches:

1. **Soft-attention** based on the work of Bahdanau et al. [1], extends the RNN-Encoder-Decoder design with an attention mechanism that allows the decoder to observe different source words for each target word.
2. **Self-attention** aims to improve on limitations of RNNs, i.e. high computational complexity and non-parallelizeable computation. For both the encoder and decoder, the Transformer model, proposed by Vaswani et al. [2], employs stacks of self-attention layers in lieu of RNNs.

Experiments & Results

We obtained a very small but fully diacritized text from the Lagos-NWU conversational speech corpus by Niekerk, et. al. We also created our own medium-sized corpus by web-crawling the two Yorùbá-language websites.

# words	Source URL	Description
24,868	rma.nwu.ac.za	Lagos-NWU corpus
50,202	theyorubablog.com	language blog
910,401	bible.com	online bible webite

Table 3: Training data subsets

To better understand the dataset split, we computed a perplexity of 575.6 for the test targets with a language model trained over the training targets. The {source, target} vocabularies for training and test sets have {11857, 18979} and {4042, 5641} word types respectively.

We built the soft-attention and self-attention models with the Python 3 implementation of OpenNMT, an open-source toolkit created by the Klein et al. Our training hardware configuration was a standard AWS EC2 p2.xlarge instance with a NVIDIA K80 GPU, 4 vCPUs and 61GB RAM.

Attention	Size	RNN	Train%	Test%
soft + dot	2L 512	LSTM	96.2	90.1
soft + add	2L 512	LSTM	95.9	90.1
soft + tanh	2L 512	GRU	96.2	89.7
soft + tanh	1L 512	GRU	97.8	89.7
self	6L 512	-	98.5	95.4

Table 4: Results

The verbs **bàjẹ́** (to spoil), or **jùlọ** (to be more than) are discontinuous morphemes, or splitting verbs. In Table 5, the first example shows the model has learnt the diacritics necessary for **lọ** following a previously predicted **jù**. In the second example, we note the ambiguity of the three occurrences of the undiacritized **si**, with two diacritized forms, **sì**, **sí**. An examination of the attention weight matrix for this example revealed that the third instance **sí** attends to the previous **sí** and that the first two attend to each other.

source	emi ni oye ju awon agba lo nitori mo gba eko re
target	èmi ni òye jù àwọn àgbà lọ nítorí mo gba ẹ̀kọ́ rẹ
prediction	èmi ni òye **jù** àwọn àgbà **lọ** nítorí mo gba ẹ̀kọ́ rẹ
source	emi yoo si si oju mi si juda
target	èmi yóò sì sí ojú mi sí ilé júdà
prediction	èmi yóò sì sí ojú mi sí ilé júdà
source	oun ko si ni ko ile ini re sile
target	òun kò sì ní **kọ** ilẹ̀ ìní rẹ̀ sílẹ̀
prediction	òun kò sì ní **kọ́** ilẹ̀ ìní rẹ̀ sílẹ̀

Table 5: Example predictions

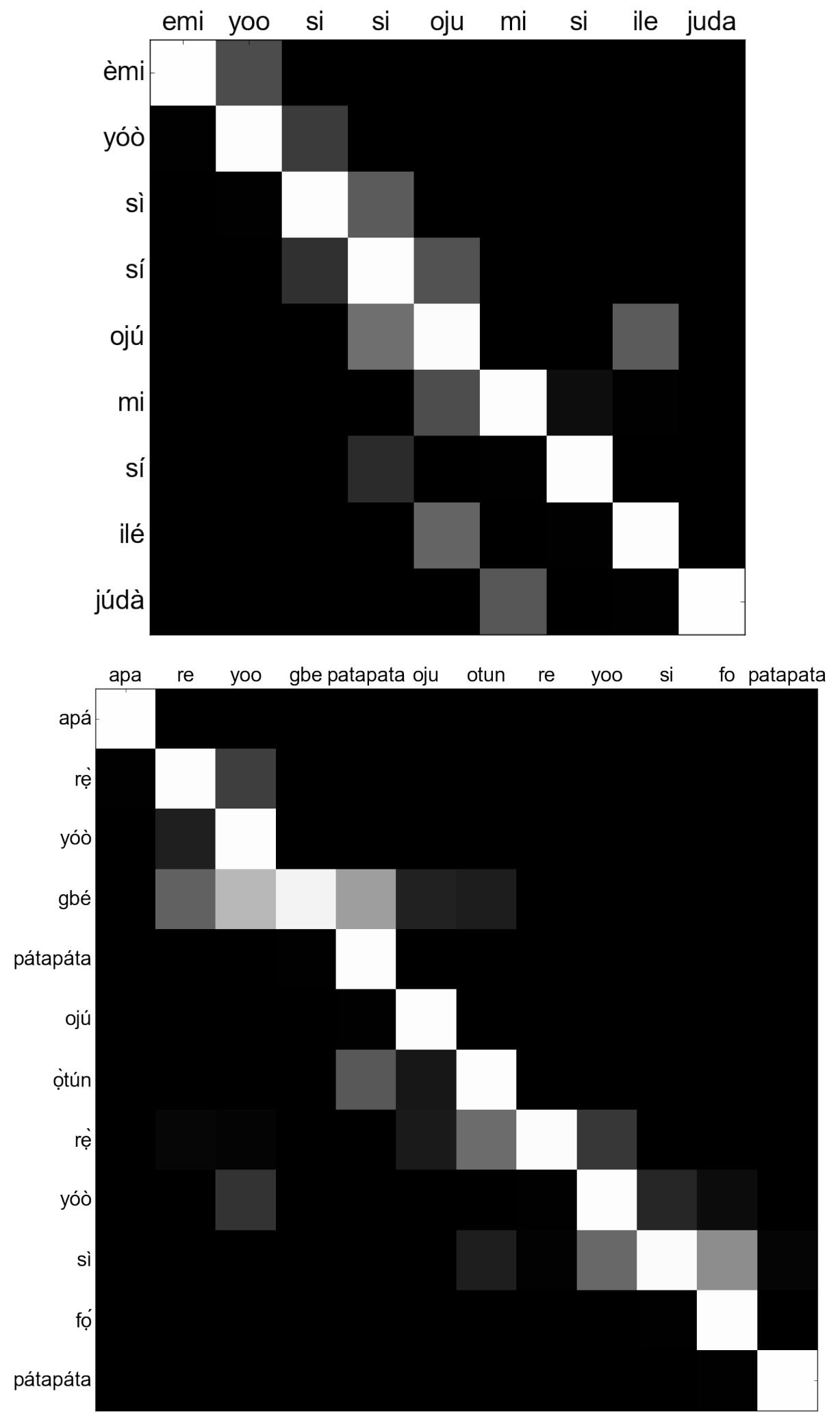

Figure 1: Attention Weights

Conclusions

- Attention-based sequence-to-sequence learning approaches perform well on the Yorùbá diacritic restoration task.
- Our approach minimizes the manual work needed to quickly create a high quality text corpus for TTS and MT tasks.
- To make ADR more suitable as a preprocessing step for end-user text processing applications or any business usage, it will be necessary to augment training the corpus with more general purpose text.

Forthcoming Research

Avenues for future work include evaluating previous approaches to Yorùbá ADR on the present dataset, growing the training corpus and training superior word embeddings. We see the application of ADR to OCR output of scanned books as a fruitful next step in building high quality text corpora.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.