Improving Yorùbá Diacritic Restoration

Anonymous authors
Paper under double-blind review

Abstract

Yorùbá is a widely spoken West African language with a writing system rich in orthographic and tonal diacritics. They provide morphological information, are crucial for lexical disambiguation, pronunciation and are vital for any computational Speech or Natural Language Processing tasks. However diacritic marks, known in Yorùbá as *amí ohùn*, are commonly excluded from electronic texts due to limited device and application support as well as general education on proper usage.

Building on our previous work (**?**), we report on recent efforts at dataset cultivation, including web-scraping, text cleaning, diacritic error correction as well as Optical Character Recognition (OCR) from books. By aggregating and improving disparate texts from the web, as well as text curated from personal libraries, we were able to significantly grow our clean Yorùbá dataset from a majority Bibilical text corpora of some 900K tokens from three sources to over 11M tokens from over a dozen sources.

We evaluate improved versions of our diacritic restoration models on a new, general purpose, public domain Yorùbá dataset of modern journalistic news text. We selected this text to be a multi-purpose corpus reflecting contemporary usage. As before, we have released all pre-trained models, datasets and source-code as an open-source project to advance efforts on Yorùbá language technology.

1 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 (**?**). The phonology is comprised of eighteen consonants *(b, d, f, g, gb, h, j, k, l, m, n, p, r, s, ṣ, t, w, y)*, seven oral vowel *(a, e, ẹ, ị, o, ọ, u)* and five nasal vowel phonemes *(an, ẹn, in, ọn, un)* with three kinds of tones realized on all vowels and syllabic nasal consonants *(ḿ, ń)* (**?**). Yorùbá orthography makes notable use of tonal diacritics to designate tonal patterns, and orthographic diacritics like underdots for various language sounds (**??**).

On modern computing and communication platforms, the majority of Yorùbá text is written in plain ASCII, without diacritics. This has negative implications for the quality of any Machine Translation (MT), Natural Language Processing (NLP) or Automatic Speech Recognition (ASR) task. To unlock the potential for a robust ecosystem of *Yorùbá-first* language technologies, Yorùbá text must be correctly represented on current and future computing environments. The ultimate objective of automatic diacritic restoration (ADR) systems is to facilitate text entry and text correction that motivates and encourages the correct orthography and promotes quotidian usage of the language in electronic media.

1.1 Ambiguity in non-diacritized text

The main challenge in non-diacritized text is that it is very ambiguous (**????**). ADR, which goes by other names such as Unicodification (**?**) or deASCIIfication (**?**), is a process which attempts to resolve the ambiguity present in undiacritized text. 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

Table 1: Diacritic characters with their non-diacritic forms

Characters		Examples
à á ǎ	**a**	gbà *(spread)*, gba *(accept)*, gbá *(hit)*
è é ẹ ẹ̀ ẹ́	**e**	èsè *(dye)*, ẹsẹ̀ *(foot)*, esé *(cat)*
ì í	**i**	ìlù *(drum)*, ilu *(opener)*, ìlú *(town)*
ò ó ọ ọ̀ ọ́ ǒ	**o**	arọ *(an invalid)*, aró *(indigo)*, àrò *(hearth)*, àrọ *(funnel)*, àrọ̀ *(catfish)*
ù ú ǔ	**u**	kùn *(to paint)*, kun *(to carve)*, kún *(be full)*
ǹ ń n̄	**n**	ǹ (a negator), n *(I)*, ń (continuous aspect marker)
ṣ	**s**	ṣà *(to choose)*, ṣá *(fade)*, sà *(to baptise)*, sá *(to run)*

applied to the vowels and syllabic nasals within the words" (**?**). Analogously to the analyses performed in (**?**), we quantify the ambiguity in our current training set 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. IOHAVOC.

Finally, 64% of all unique, non-diacritized monosyllabic words possess multiple diacritized forms (**??**). Given that Yorùbá verbs are predominantly monosyllabic and that there are tonal changes rules governing how tonal diacritics on a specific word are *altered* based on context, we acknowledge the complexity of the disambiguation task of diacritic restoration (**??**).

1.2 Improving generalization performance

In our efforts to make the the first ADR models (**?**), trained on majority Biblical text, available to a wider audience, we frequently tested on colloquial, conversational text. We observed that these early ADR models suffered from domain-mismatch generalization errors and appeared particularly not-robust when presented with contractions or variants of common phrases. We attributed these errors to low-diversity of sources, in this case just three, as well as not-enough-data, under a million tokens. To remedy this problem, we set about to we demonstrate that we can improve the model trained on majority Biblical text, using a lot more text from the a variety of sources.

This paper is organized as follows. Section 2 summarizes our data collection efforts. In section 3, we outline our experimental setup and models trained. In section 4, we present the results of the evaluation. In section 5, we conclude with an error analysis and future directions.

2 Data Collection

We started by first attempting a comprehensive GitHub, literature and Google web search, assembling all the public-domain Yorùbá sources online. Those without admissible or consistent quality were put into a special queue for human supervision and corrections. This notably included Wikipedia and Twitter. Below we describe briefly each of the admitted corpora to convey a sense of the subject matter and word distributions.

- **JW300**: JW300 is a large-scale a parallel corpus for Machine Translation (MT) comprising more than three hundred languages with on average one hundred thousand parallel sentences per language pair. The Yorùbá-English token pairs number some thirteen million. JW300 text is drawn from a number of online blogs, news and contemporary religious magazines by Jehovah's Witnesses (JW).
- **Bíbélì**: Two different versions of the Yorùbá Bible were web-scraped cleaned. The first is from Biblica, which is a translation of the New International Version (NIV). The second version is a 2010 translation published by the Bible Society of Nigeria (BSN). Our previous work only used the Biblica version. A computational analysis of the word distributions of each Bible's text revealed that on average 60% of all verses were not identical. Therefore,

it valuable to have both versions of the Bible to give a diversity of expression for the same concepts. Further more, the additional size was is still dwarfed by the JW300 corpus.

- **Language identification corpus**: This medium sized corpus of one hundred and fifty thousand tokens was used in a Nigerian language identification task involving the three major Nigerian tongues, Hausa, Igbo and Yorùbá. We obtained the public-domain text from GitHub **?**.
- **Yorùbá-Twi word-embedding corpora** This medium sized corpus comprises a number of texts used in a recent word embedding task
- **Òwe**: Òwe is a small collection of Yorùbá proverbs scraped from the electronic version of *The Good Person: Excerpts from the Yorùbá Proverb Treasury*, a collection compiled and translated by Dr. Oyekan Owomoyela at the University of Nebraska - Lincoln. The corpus has five thousand tokens and parallel English translations **?**.
- **Universal Declaration of Human Rights** This is a tiny corpus of some two thousand eight hundred tokens from the Universal Declaration of Human Rights.
- **Lẹ́síkà**: ẹ́síkà is a fifty thousand token corpus of unigrams taken from recent research by Victor Williamson on the Yorùbá dictionary. To this list we added de-duped unigrams from all the entire dataset, before splitting into training and validation. This gave us a comprehensive view and the ability to learn the diacritic patterns of all Yorùbá words as well as common loan-words.
- **Private texts**: In addition to the public domain texts, the authors also had access to personal archives of written work, correctly tone-marked that would add some meaningful diversity to the corpus. These included transcription of interviews and long form stories. These texts numbered some three hundred thousand tokens.
- **Yorùbá Object Character Recognition (OCR)**: Another direction explored by the authors, was the use of Object Character Recognition (OCR). This is the process of scanning physical books to derive plain text from the scanned images. However the presence of diacritics in Yorùbá books presents problems for English OCR models. So we retrained a new Yorùbá OCR model using existing clean text. TIMI elborate here on what was done. Feel free to blow grammer and give technical details.Ẇe bootstrapped our Háà Ènìyàn efforts, Ogboju Ode as well as Aaro Meta.

Table 2: Training data subsets

# words	Source or URL	Description
24,868	rma.nwu.ac.za	Lagos-NWU corpus
50,202	theyorubablog.com	language blog
910,401	bible.com/versions/911	Biblica
11,488,825	opus.nlpl.eu	JW300
831,820	bible.com/versions/207	Bible Society Nigeria
142,991	-	Language ID corpus
47,195	-	Yorùbá Lexicon
29,338	yoruba.unl.edu	Proverbs
28,308	yo.globalvoices.org	Global Voices news
2,887	unicode.org/udhr	Human rights edict
150,360	-	Interview text
15,243	-	Short Stories
910,401	-	Háà Ènìyàn

3 Methodology and Results

3.1 Community Building

Online communication. The community channel that gained most participation was Slack. Figure **??** shows how the number of members grew since creation. Of the weekly active members, roughly one half submit posts. End of November, something magic happened that doubled the number of members. We can also see that the upcoming paper deadlines increased activity.

3.2 Focus

3.3 Discoverability

3.4 Reproducibility

Transparency. Not only the code, data and results, but also the meeting notes and discussions are publicly available, such that any one with interest can re-trace decisions, re-read discussions, and re-produce benchmarks. In this way, new members to the community can catch up and find their areas of interest without the need of a central director.

Minimal assumptions of resources. The goal is to build data and code that anyone should be able to reproduce. This means that we cannot afford to make assumptions about access to hardware, to specialized knowledge in NLP, or to local experts. Therefore, created resources and their usage are designed to be self-explanatory. This is facilitated largely through Jupyter Notebooks[1] comprising documented data creation, model configuration, training and evaluation, optimized to run on Google Colab with a single (free) GPU for a small limited number of hours. The NMT models are built on Joey NMT (**?**), an NMT toolkit that comes with a beginner-friendly documentation.

3.5 Benchmark Creation

Global test set. For transfer learning it has to be guaranteed that there is no overlap between training data of any language with test data of any other language. For that reason we chose the JW300 corpus (**?**), that covers 300 languages of which 169 were identified as African, to extract a set of English sentences that are regarded as global test set and are excluded from training data for any language. Of the 169 languages, 101 were accessible with the `opusTools` package.[2] Figure **??** summarizes the size of the parallel data by language. It ranges from 1,784 sentences for Mende, to 1.1M for Afrikaans, thus covers a wide range of what is understood as low-resource. From those sentences, we choose 4,000 English sentences that are shared across the most languages and are longer than eight words. This results in test sets of varying size per language. Possible expansions of these test sets (filling up smaller sets, adding data from different domain, etc.) will require an update of the global test set.

3.6 Addressing Decentralization

3.7 Language Complexity

4 Preliminary Results

4.1 Discussion

4.2 Future Work

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

[1] https://jupyter.org/

[2] https://github.com/Helsinki-NLP/OpusTools