IMPROVING YORÙBÁ DIACRITIC RESTORATION

David Adélaní
Saarland University
Niger-Volta LTI

Iroro Fred Ọ̀nọ̀mẹ̀ Orife
Niger-Volta LTI

Kọ́lá Túbọ̀sún
Yorùbá Name

Tìmílẹ́hìn Fasubaa
Niger-Volta LTI

Ọlámilékan Wahab
Niger-Volta LTI

Wúràọ̀lá Oyèwùsì
Data Science Nigeria
Niger-Volta LTI

Victor Williamson
Yorùbá Name

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 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. By aggregating and improving disparate texts from the web and various personal libraries, we were able to significantly grow our clean Yorùbá dataset from a majority Bibilical text corpora from three sources to millions of tokens from over a dozen sources. We evaluate updated diacritic restoration models on a new, general purpose, public domain Yorùbá dataset of modern journalistic news text, selected to be a multi-purpose corpus reflecting contemporary usage. All pre-trained models, datasets and source-code have been released 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 (Wikipedia, 2004). 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)* (Akinlabi, 2004). Yorùbá orthography makes notable use of tonal diacritics, known as *amí ohùn*, to designate tonal patterns, and orthographic diacritics like underdots for various language sounds (Adegbola & Odilinye, 2012; Wells, 2000).

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 (Orife, 2018; Adegbola & Odilinye, 2012; Asahiah et al., 2017; De Pauw et al., 2007). ADR, which goes by other names such as Unicodification (Scannell, 2011) or deASCIIfication (Arslan, 2016), is a process which attempts to resolve the ambiguity present in undiacritized text. Adegbola et al. state that for ADR

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**	aro *(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)*

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" (Adegbola & Odilinye, 2012). Analogously to the analyses performed in (Orife, 2018), 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.

Finally, 64% of all unique, non-diacritized monosyllabic words possess multiple diacritized forms (Oluseye, 2003; Delano, 1969). 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 (Orife, 2018; Delano, 1969).

1.2 Improving generalization performance

In our efforts to make the the first ADR models (Orife, 2018), 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 Methodology

2.1 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), by Jehovah's Witnesses (JW), 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 (Agić & Vulić, 2019).

- **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 (Asubiaro et al., 2018).
- **Yorùbá-Twi word-embedding corpora** This medium sized corpus comprises a number of texts used in a recent word embedding task
- **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 (Owomoyela, 2003).
- **Universal Declaration of Human Rights** is a tiny corpus of some two thousand eight hundred tokens from the Universal Declaration of Human Rights.
- **Lẹ́síkà** is a fifty thousand token corpus of unigrams taken from recent research by Victor Ẇilliamson 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 H nyn 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	-	H nyn

2.2 EXPERIMENTAL SETUP

When collecting data from disparate sources, we preprocessed texts to ensure consistent, error-free diacritization, splitting lines on full-stops to give one sentence per line. To ensure our splits are drawn from similar distributions, we combined all text, shuffled and split utterances into a ratio of 90%, 10%, for training and dev sets respectively.

To prepare source and target texts for parallel training via a sequence-to-sequence architecture, all characters in the texts were dispossessed of their diacritics. Diacritized text was converted to Unicode Normalization Form Canonical Decomposition (NFD) which separates a base character from its diacritics. Next, *UnicodeCategory.NonSpacingMark* characters, which house the diacritic modifications to a character, were filtered out. This yielded two sets of text, one stripped of diacritics (the source) and the other with full diacritics (the training target).

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 (Stolcke, 2002). The {source, target} vocabularies for training and test sets have {11857, 18979} and {4042, 5641} word types respectively.

2.3 USING PRE-TRAINED TEXT EMBEDDINGS

TEXT TO FIX and TO DESCRIBE DAVID's FastText, BERT and XLM embeddings. FastText, BERT and XLM embeddingsFastText, BERT and XLM embeddingsFastText, BERT and XLM embeddingsFastText, BERT and XLM embeddingsFastText, BERT and XLM embeddingsFastText, BERT and XLM embeddingsFastText, BERT and XLM embeddings.

2.4 A NEW, MODERN MULTI-PURPOSE EVALUATION DATASET

Here we discuss in detail the selection of Iroyin, perhaps present some facts and figures about it, to better understand why it is a good evaluation set and to motivate how we converged on it being the best text subset. Modern, non-archaic text, popular usage and a better proxy than any of the religious centered texts (JW300, Bibeli, etc) and less formal.

Global Voices is a multilingual community of journalists, translators, bloggers, academics and human rights activists. Their newsroom articles are translated into dozens of languages. We used a web-scrape of their Yorùbá articles comprising some twenty-eight thousand tokens.

In the previous paper, the test set was selected as a subset of the total available dataset, this means that the test and validation sets, drew generally speaking from the same distribution as the training set, this explains the stellar performance reported in Orife, 2018. However, we need our evaluation set to be a (1) public (2) representative of a mixture of modern written and spoken Yoruba styles (3) large enough.

From the texts that we had available, including interviews, journalistic news, literary and liturgical texts, there was a consensus that the journalistic news text best represented the modern, colloquial usage of the language. Therefore, by evaluating the quality of the various models on this diacritic restoration task would ensure that published models would best match the language usage expectations of users. For these reasons we converged on using broadcast news (Iroyin) as our evaluation set.

2.5 TRAINING

3 RESULTS

The dev set will be the test set drawn from the training data's distribution. Test set will be Iroyin as described above. We should match the figures published in (Orife, 2018) with the current figure to tell a consistent story. Even though the test set has changed, we should still evaluate the legacy models on it to show how badly it geneneralized even thought it had very good in-domain performance.

To evaluate the performance of our ADR models, we computed the accuracy score as the ratio of correct words restored to all words. We calculate the perplexity of each model's predictions based

on the test set targets. Discussion of results. Discussion of results. Discussion of results. Discussion

Table 3: Training & Test Accuracy and Perplexity

Model	Accuracy %	Perplexity
Baseline RNN (?)	90.1	1.68
Baseline Bandahau	90.1	1.85
Bandahau+	90.1	1.9
Transformer+	90.1	1.9
Bandahau+JW300	90.1	1.9
Transformer+JW300	90.1	1.9
Bandahau+JW300+FastText	90.1	1.9
Transformer+JW300+FastText	90.1	1.9

of results. Discussion of results. Discussion of results. Discussion of results. Discussion of results. Discussion of results. Discussion of results.

3.1 Error Analysis

While performing error analyses on the model predictions, we observed: DESCRIBE YOUR OBSERVATIONS

4 Conclusions

4.1 Discussion

Additional data and more diverse data definitely improves performance. Modern Text embeddings provide an additional boost in accuracy (TBD).

4.2 Future Work

All public-domain datasets referenced in this work are available on GitHub.[1] [2]

Acknowledgments

We thank everyone from YorubaName, Masakhane, DataScience Nigeria. Who else are we thanking?

References

Tunde Adegbola and Lydia Uchechukwu Odilinye. Quantifying the effect of corpus size on the quality of automatic diacritization of Yorùbá texts. In *Spoken Language Technologies for Under-Resourced Languages*, pp. 48–53, 2012.

Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1310. URL https://www.aclweb.org/anthology/P19-1310.

Akinbiyi Akinlabi. The sound system of Yorùbá. *Lawal, N. Sadisu, MNO & Dopamu, A (Eds.) Understanding Yoruba life and culture. Trento: Africa World Press Inc*, pp. 453–468, 2004.

Ahmet Arslan. Deasciification approach to handle diacritics in Turkish information retrieval. *Information Processing & Management*, 52(2):326–339, 2016.

[1] https://github.com/Niger-Volta-LTI/yoruba-text

[2] https://github.com/Niger-Volta-LTI/yoruba-adr

Franklin Oladiipo Asahiah, Odetunji Ajadi Odejobi, and Emmanuel Rotimi Adagunodo. Restoring tone-marks in standard Yorùbá electronic text: improved model. *Computer Science*, 18(3):301–315, 2017. ISSN 2300-7036. URL https://journals.agh.edu.pl/csci/article/view/2128.

Toluwase Asubiaro, Tunde Adegbola, Robert Mercer, and Isola Ajiferuke. A word-level language identification strategy for resource-scarce languages. In *Proceedings of the Association for Information Science and Technology*, volume 55, 11 2018. doi: 10.1002/pra2.2018.14505501004.

Guy De Pauw, Peter W Wagacha, and Gilles-Maurice De Schryver. Automatic diacritic restoration for resource-scarce languages. In *International Conference on Text, Speech and Dialogue*, pp. 170–179. Springer, 2007.

Isaac O Delano. *A dictionary of Yorùbá monosyllabic verbs*, volume 1. Institute of African Studies, University of Ife, 1969.

Adesola Oluseye. Yorùbá: A Grammar Sketch: Version 1.0, 2003. URL https://www.bible.com/bible/911/GEN.1.BMY.

Iroro Fred Ọ̀nọ̀mẹ̀ Orife. Sequence-to-Sequence Learning for Automatic Yorùbá Diacritic Restoration. In *Proceedings of the Interspeech*, pp. 27–35, 2018.

Oyekan Owomoyela. The good person: Excerpts from the Yorùbá proverb treasury, 2003. URL http://yoruba.unl.edu.

Kevin P Scannell. Statistical unicodification of African languages. *Language resources and evaluation*, 45(3):375, 2011.

Andreas Stolcke. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*, 2002.

JC Wells. Orthographic diacritics and multilingual computing. *Language problems and language planning*, 24(3):249–272, 2000.

Wikipedia. Yoruba language, wikipedia, June 2004. URL https://en.wikipedia.org/wiki/Yoruba_language.

A APPENDIX

You may include other additional results here.

Table 4: Training & Test Accuracy and Perplexity

Model	**Train%**	**Dev%**	**Test%**	**PPL**
Baseline RNN	96.2	90.1	90.1	1.68
Bandahau from [1]	95.9	90.1	90.1	1.85
Bandahau++	-	-	-	-
Transformer++	-	-	-	-
Transformer++ FastText	-	-	-	-