IMPROVING YORÙBÁ DIACRITIC RESTORATION

Anonymous authors
Paper under double-blind review

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. The phonology is comprised of eighteen consonants, seven oral vowel and five nasal vowel phonemes with three kinds of tones realized on all vowels and syllabic nasal consonants (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).

Diacritics provide morphological information, are crucial for lexical disambiguation, pronunciation and are vital for any computational Speech or Natural Language Processing 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 attempts to decode the ambiguity present in undiacritized text. Adegbola et al. assert 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" (Adegbola & Odilinye, 2012).

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)*

1.2 IMPROVING GENERALIZATION PERFORMANCE

In our efforts to make the the first open-sourced ADR models available to a wider audience, we frequently tested on colloquial, conversational text. We observed that these early models suffered from domain-mismatch generalization errors and appeared particularly not-robust when presented with contractions or variants of common phrases. Because they were trained on majority Biblical text, we attributed these errors to low-diversity of sources and an insufficient number of training examples. To remedy this problem, we set about to we demonstrate that we can improve the model trained on majority Biblical text, using appreciably more text from a variety of sources.

2 METHODOLOGY

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. TIMI OCR SENTENCE HERE. In Table 2 we summarize each of the admitted corpora to convey a sense of the subject matter and word distributions.

Table 2: Data sources

# 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
2,887	unicode.org/udhr	Human rights edict
—	YorubaTWI	YorubaTwi Embeddings
150,360	Private sources	Interview text
15,243	OCR	Short Stories
910,401	OCR	Háà Èníyàn
28,308	yo.globalvoices.org	Global Voices news

Experimental setup Data preprocessing and parallel text preparation followed the same procedures used in (Orife, 2018). Experiments included evaluations of the effect of the various texts on the test accuracy, notably for JW300, which is a disproportionately large contributor to the dataset. We also evaluated models trained with pre-trained FastText embeddings to understand the boost in performance possible with word embeddings.

A new, modern multi-purpose evaluation dataset To succeed at the ADR task for users, our research experiments needed to be guided by a test set based around modern, non-archaic text in popular usage. After much review, there was a consensus that the journalistic news text best represented the modern, colloquial usage of the language. So we selected Global Voices, 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.

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

4 CONCLUSIONS AND FUTURE WORK

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

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

ACKNOWLEDGMENTS

We thank everyone from Yorùbá Name, Masakhane, DataScience Nigeria.

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.

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.

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.

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.

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

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

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

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

A APPENDIX

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

Table 4: Example Sentences

Source:	Still , words of apology are a strong force toward making peace
Reference:	We can even ask God to create in us a pure heart
Prediction:	We can even ask God to create in us a pure heart
Source:	Still , words of apology are a strong force toward making peace
Reference:	We can even ask God to create in us a pure heart
Prediction:	We can even ask God to create in us a pure heart
Source:	Still , words of apology are a strong force toward making peace
Reference:	We can even ask God to create in us a pure heart
Prediction:	We can even ask God to create in us a pure heart