IMPROVING YORÙBÁ DIACRITIC RESTORATION

Iroro Fred Ọ̀nọ̀mẹ̀ Orife
Niger-Volta LTI

David I. Adélaní
Saarland University, Niger-Volta LTI

Timi Fasubaa
Niger-Volta LTI

Victor Williamson
University of Wisconsin-Milwaukee

Wuraola Fisayo Oyewusi
Data Science Nigeria

Ọlámilékan Wahab
Niger-Volta LTI

Kọ́lá Túbọ̀sún
Yorùbá Name

iroro@alumni.cmu.edu, didelani@lsv.uni-saarland.de
timifasubaa@berkeley.edu, victorlamont05@gmail.com
oyewusiwuraola@gmail.com, olamyy53@gmail.com, kolatubosun@gmail.com

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 and pronunciation, and are vital for any computational Speech or Natural Language Processing (NLP) task. To build a robust ecosystem of *Yorùbá-first* language technologies, Yorùbá text must be correctly represented in computing environments. The ultimate objective of automatic diacritic restoration (ADR) systems is to facilitate text entry and text correction that 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; Asahiah et al., 2017; Adegbola & Odilinye, 2012; 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**	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)*

1.2 IMPROVING GENERALIZATION PERFORMANCE

To make the first open-sourced ADR models available to a wider audience, we tested extensively on colloquial and conversational text. These soft-attention seq2seq models (Orife, 2018), trained on the first three sources in Table 2, suffered from domain-mismatch generalization errors and appeared particularly weak when presented with contractions, loan words 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 aggregated text from a variety of online public-domain sources as well as actual books. After scanning physical books from personal libraries, we successfully employed commercial Optical Character Recognition (OCR) software to concurrently use English, Romanian and Vietnamese characters, forming an *approximative superset* of the Yorùbá character set. Text with inconsistent quality was put into a special queue for subsequent human supervision and manual correction. The post-OCR correction of Háà Ènìyàn, a work of fiction of some 20,038 words, took a single expert two weeks of part-time work by to review and correct. Overall, the new data sources comprised varied text from conversational, various literary and religious sources as well as news magazines, a book of proverbs and a Human Rights declaration.

2 METHODOLOGY

Experimental setup Data preprocessing, parallel text preparation and training hyper-parameters are the same as in (Orife, 2018). Experiments included evaluations of the effect of the various texts, 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 (Alabi et al., 2020; Bojanowski et al., 2017). Our training hardware configuration was an AWS EC2 p3.2xlarge instance with `OpenNMT-py` (Klein et al., 2017).

Table 2: Data sources, prevalence and category of text

# 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 (NIV)
11,488,825	opus.nlpl.eu	JW300
831,820	bible.com/versions/207	Bible Society Nigeria (KJV)
177,675	GitHub	Embeddings dataset (mixed)
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	Conversational interviews
15,281	Private sources	Short stories
20,038	OCR	Háà Ènìyàn (Fiction)
28,308	yo.globalvoices.org	Global Voices news

A new, modern multi-purpose evaluation dataset To make ADR productive for users, our research experiments needed to be guided by a test set based around modern, colloquial and not exclusively literary text. After much review, we selected Global Voices, a corpus of journalistic news text from a multilingual community of journalists, translators, bloggers, academics and human rights activists (Global Voices, 2005).

3 RESULTS

We evaluated the ADR models by computing a single-reference BLEU score using the Moses `multi-bleu.perl` scoring script, the predicted perplexity of the model's own predictions and the Word Error Rate (WER). All models with additional data improved over the 3-corpus soft-attention baseline, with JW300 providing a {33%, 11%} boost in BLEU and absolute WER respectively. Error analyses revealed that the Transformer was robust to receiving digits, rare or code-switched words as input and degraded ADR performance gracefully. In many cases, this meant the model predicted the undiacritized word form or a related word from the context, but continued to correctly predict subsequent words in the sequence. The FastText embedding provided a small boost in performance for the Transformer, but was mixed across metrics for the soft-attention models.

Table 3: BLEU, 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**

4 CONCLUSIONS AND FUTURE WORK

Promising next steps include further automation of our human-in-the-middle data-cleaning tools, further research on contextualized word embeddings for Yorùbá and serving or deploying the improved ADR models[1][2] in user-facing applications and devices.

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.

Jesujoba O Alabi, Kwabena Amponsah-Kaakyire, David I Adelani, and Cristina España-Bonet. Massive vs. Curated Word Embeddings for Low-Resourced Languages. The Case of Yorùbá and Twi. In *LREC*, 2020.

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`.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017. ISSN 2307-387X.

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

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

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.

Stichting Global Voices. Global Voices. `https://yo.globalvoices.org`, 2005. Accessed: 2020-02-12.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proc. ACL*, 2017. doi: 10.18653/v1/P17-4012. URL `https://doi.org/10.18653/v1/P17-4012`.

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.

A Appendix

Table 4: The best performing Transformer model trained with the FastText embedding was used to generate predictions. The Baseline model is the 3-corpus soft-attention model. ADR errors are in **red**, robust predictions of rare, loan words or digits in **green**.

Source:	mo ro o wipe awon obirin ti o ronu lati se ise ti okunrin maa n se gbodo gberaga .
Reference:	mo rò ó wípé àwọn obìrin tí ó ronú láti ṣe iṣẹ́ tí ọkùnrin máa ń ṣe gbọ̀dọ̀ gbéraga .
Prediction:	mo rò ó wípé àwọn obìrin tí ó ronú láti ṣe iṣẹ́ tí ọkùnrin máa ń ṣe gbọ̀dọ̀ gbéraga .
Baseline:	mo rò ó wípé àwọn obìrin tí ó ronú láti ṣe iṣẹ́ tí ọkùnrin máa **darí sọ̀rọ̀ láṣe lọ́kan** .
Source:	bi o tile je pe egbeegberun ti pada sile .
Reference:	bí ó tilẹ̀ jẹ́ pé ẹgbẹẹgbẹ̀rún ti padà sílé .
Prediction:	bí ó tilẹ̀ jẹ́ pé ẹgbẹẹgbẹ̀rún ti padà sílé .
Baseline:	bí ó tilẹ̀ jẹ́ pé ẹgbẹẹgbẹ̀rún tí padà **sílẹ̀ sọ̀rọ̀**
Source:	mo awon ondije si ipo aare naijiria odun **2019**
Reference:	mọ àwọn òǹdíje sí ipò àárẹ nàìjíríà ọdún **2019**
Prediction:	mọ àwọn **ondije** sí ipò àárẹ nàìjíríà ọdún **2019**
Baseline:	mo àwọn **ojojúmọ́** sí ipò **àárẹ̀** nà**í**jíríà ọdún **kiki**
Source:	iriri akobuloogu **zone9** ilu ethiopia je apeere .
Reference:	ìrírí **akọbúlọ́ọ̀gù zone9** ìlú ethiopia jẹ́ àpẹẹrẹ .
Prediction:	ìrírí **akobuloogu orílè** ìlú ethiopia jẹ́ àpẹẹrẹ .
Baseline:	ìrírí **àwọn** ìlú **esinsin arákùnrin** jẹ́ àpẹẹrẹ .
Source:	alaga akoko ilu-ti-ko-fi-oba-je ti china mao zedong ti yo awon eniyan ninu isoro .
Reference:	alága àkọ́kọ́ **ìlú-tí-kò-fi-ọba-jẹ** ti **china mao zedong** ti yọ àwọn ènìyàn nínú ìṣòro .
Prediction:	alága àk**ókò ilu-ti-ko-fi-oba-je** ti **china mao tse** ti yọ àwọn ènìyàn nínú ìṣòro .
Baseline:	**jéhóṣáfátì àkókò samáríà** tí china **lẹṣẹkẹṣẹ apá tí wà atí** ènìyàn nínú ìṣòro .