Improving Yor\`ub\'a Diacritic Restoration

Abstract

Yor\`ub\'a 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. 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 Yor\`ub\'a dataset from a majority Bibilical text corpora with three sources to millions of tokens from over a dozen sources. We evaluate updated diacritic restoration models on a new, general purpose, public-domain Yor\`ub\'a evaluation dataset of modern journalistic news text, selected to be multi-purpose and reflecting contemporary usage. All pre-trained models, datasets and source-code have been released as an open-source project to advance efforts on Yor\`ub\'a 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. 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 BIBREF0. 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 BIBREF1, BIBREF2.

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.

Introduction ::: Ambiguity in non-diacritized text

The main challenge in non-diacritized text is that it is very ambiguous BIBREF3, BIBREF4, BIBREF1, BIBREF5. 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" BIBREF1.

Introduction ::: 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 BIBREF3, trained on the first three sources in Table TABREF5, 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.

Methodology ::: Experimental setup

Data preprocessing, parallel text preparation and training hyper-parameters are the same as in BIBREF3. 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 BIBREF6, BIBREF7. Our training hardware configuration was an AWS EC2 p3.2xlarge instance with OpenNMT-py BIBREF8.

Methodology ::: 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 BIBREF9.

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.

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 in user-facing applications and devices.