TOWARDS NEURAL MACHINE TRANSLATION FOR EDOID LANGUAGES

Iroro Fred Ọ̀nọ̀mẹ̀ Orife
Niger-Volta Language Technologies Institute
Seattle, WA 98119, USA
iroro@alumni.cmu.edu

Ji Q. Ren & Yevgeny LeNet
Department of Computational Neuroscience
University of the Witwatersrand
Joburg, South Africa
{robot,net}@wits.ac.za

ABSTRACT

Many Nigerian languages have lost their previous prestige and purpose in modern society due the status of English and Nigerian Pidgin. These language inequalities for L1 speakers manifest themselves in unequal access to information, connectivity, health care, security as well as attenuated participation in political and civic life. This work explores the feasibility of Neural Machine Translation (NMT) for Edoid Languages, spoken by some 5 million people in Southern Nigeria. Using public datasets, we trained and evaluated translation models for four widely spoken langauges in this family: Ẹ̀dó, Ẹ̀sán, Urhobo and Isoko. Trained models, code and datasets have also been open-sourced to advance future research efforts on Edoid language technology.

1 INTRODUCTION

Belonging to the Volta-Niger family, Edoid languages are a group some two dozen languages spoken in southern Nigeria by about 5 million people. The term *Edoid* comes from Ẹ̀dó primary language of the famed Kingdom of Benin and the most broadly spoken member.

Good Goveranance, language equality, access to information and the such.

1.1 LANGUAGES

1.2 RELATED WORKS

While there has been recent interest in NMT for African languages, in Nigeria there has been a bit of literature on Rule-based, phrase-based and Statistical machine translation. This is the first work known to the authors done in any of the Edoid langauges specifically for machine translation.

2 METHODOLOGY

2.1 DATASET

The recently released JW300 dataset is a large-scale, parallel corpus for Machine Translation (MT) comprising more than three hundred languages with on average one hundred thousand parallel sentences per language pair. English-{Ẹ̀dó, Ẹ̀sán, Urhobo, Isoko} token pairs number {10200, 2000, 200, 4000} respectively. JW300 text is drawn from a number of online blogs, news and contemporary religious magazines by Jehovah's Witnesses (JW).

2.2 MODELS

We used the JoeyNMT framework to train the Transformer.

3 RESULTS

Table 1: 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	-	-	-	-
Transformer++ BERT	-	-	-	-
Transformer++ XLM	-	-	-	-

more discussion here

3.1 HEADINGS: SECOND LEVEL

Second level headings are in small caps, flush left and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

3.2 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 will also provide an additional boost in accuracy. Overall more studies are needed regarding algorithmic preprocessing and hyperparamter fine-tuning. For example, we naively saw that for the smaller corpora BPE tokenization gave a slight boost in BLEU performance, while

4.2 FUTURE WORK

We see this work as a foundational effort on a few fronts. Thee include social justice by addressing an aspect of technological language inequality, language perservation and by establishing baselines and from which to build on. Given the comparatively low (Oladele Awobuluyi) litearay traditions but the very strong oral traditions, foundational language technologies based on good clean text, like language and translation models are just the start, but very important precusor to speech interfaces. Imagine a world in which a culture rooted in a strong oral tradition can make use of Speech-to-Speech interfaces, speaking and being spoken to idiomatically. This is where the future of African langauge technology lies and mahcine translation and good clean datasets are the core.

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

ACKNOWLEDGMENTS

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

A APPENDIX

You may include other additional sections here.

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