TOWARDS NEURAL MACHINE TRANSLATION FOR EDOID LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Many Nigerian languages have relinquished their previous prestige and purpose in modern society to English and Nigerian Pidgin. For the millions of L1 speakers of indigenous languages, there are inequalities that manifest themselves as unequal access to information, communications, health care, security as well as attenuated participation in political and civic life. To minimize exclusion and provide more access, this work explores the feasibility of Neural Machine Translation (NMT) for the Edoid language family of Southern Nigeria. Using the new JW300 public dataset, we trained and evaluated baseline translation models for four widely spoken languages in this group: Ẹ̀dó, Ẹ̀sán, Urhobo and Isoko. Trained models, code and datasets have been open-sourced to advance future research efforts on Edoid language technology.

1 INTRODUCTION

Nigerian languages with primarily oral traditions have been the most susceptible to language endangerment. Numerous have already ceded privileged positions in society to English or Nigerian Pidgin. Unlike many East and South Asian societies, which preserved the socio-linguistic status of their indigenous languages for centuries under colonial rule, Nigerian tongues have not fared as well. For L1 speakers, this language inequality is further exacerbated in a technological age, where only the most highly resourced (i.e. colonial) languages become the milieu for access to information, telecommunications, health care, security, economic advancement as well as participation in political and civic life Odoje (2013); Awobuluyi (2016). There are practical and technical challenges with respect to orthographic standardizations, consistent diacritic representation (Unicode) in electronic media and across device types.

For almost-extinct languages, machine translation offers technology for language documentation and preservation. For speakers of minority Nigerian languages, it can facilitate good governance, national development and offers a path for technological, economic, social and political participation and empowerment to those with unequal access (Odoje, 2016; 2013).

The objective of this paper is to report foundational NMT work to assist translators and the layperson alike working in Edoid languages. We hope to bootstrap the development and sustenance of scholarly and literary traditions, beyond religious texts. By making our models and code broadly accessible as open-source projects, we hope to energize academic and industry interest while contributing tools of self- solution of disenfranchisement based on language and socio-economic barriers.

1.1 LANGUAGES

Belonging to the Volta-Niger family, and spoken by some 5 million people, the Edoid languages of Southern Nigeria (Edo and Delta State) comprise over two dozen so-called “minority” languages. The term *Edoid* stems from Ẹ̀dó, the most broadly spoken member langauge and the language of the famed Kingdom of Benin. Urhobo and Isoko are classified as South-Western Edoid, while Ẹ̀dó, Ẹ̀sán are classified as North-Central (Wikipedia contributors, 2005).

All these languages and tonal in nature, ie. have certain grammatical and tonal relationships, blah blah blah

We contrast the performance of a baseline Transformer model across the four languages under study, examining the effect of word-level versus subword-level tokenization.

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 of which 101 are African (Agić & Vulić, 2019). English-{Ẹ̀dó, Ẹ̀sán, Urhobo, Isoko} token pairs cardinality is itemized in Table 3. {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. We built all models with the Python 3 implementation of `JoeyNMT`, an open-source toolkit created by the Klein et al. (**?**). Our training hardware configuration was a standard AWS EC2 p2.xlarge instance with a NVIDIA K80 GPU, 4 vCPUs and 61GB RAM. Training the various models took place over the course of a few days.

3 Results

Language	**BPE**		**Word**		**Training Tokens**
	dev	test	dev	test	
Ẹ̀dó	7.92	12.49	5.99	8.24	229,307
Ẹ̀sán	4.94	6.25	3.39	5.30	87,025
Urhobo	15.91	28.82	11.80	22.39	519,981
Isoko	32.58	38.05	32.38	38.91	4,824,998

3.1 Qualitative

Unsuprisingly, for Urhobo and Isoko which are much better resourced, the BLEU scores are generally correlated with the translation quality when reviewed by L1 speakers. For example, for Urhobo this translation captures much o

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

The authors thank Dr. Ajovi B. Scott-Emuakpor, MD and Dr. John Nevboyeri Orife for their encouragement, support and qualitative translations.

REFERENCES

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.

Ọládélé Awobuluyi. Why We Should Develop Nigerian Languages. *Issues in Contemporary African Linguistics: A Festschrift for Oladele Awobuluyi*, 11:347, 2016.

Clement Odoje. Language Inequality: Machine Translation as the Bridging Bridge for African languages. 4, 01 2013.

Clement Odoje. The Peculiar Challenges of SMT to African Languages. *ICT, Globalisation and the Study of Languages and Linguistics in Africa*, pp. 223, 2016.

Wikipedia contributors. Edoid languages, 2005. URL https://en.wikipedia.org/wiki/Edoid_languages. [Online; accessed 2020-01-30].

A APPENDIX

You may include other additional sections here.

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