

TOWARDS NEURAL MACHINE TRANSLATION FOR EDOID LANGUAGES

Anonymous authors

Paper under double-blind review

1 INTRODUCTION

Many of the over 500 languages are spoken in Nigeria today have relinquished their previous prestige and purpose in modern society to English and Nigerian Pidgin, notably amongst the younger generations. Unlike numerous East and South Asian societies, which preserved the socio-linguistic status of their indigenous languages under colonial rule, Nigerian communities with primarily oral traditions have been the most susceptible to language endangerment (Rolle, 2013; Omo-Ojugo, 2004).

For tens of millions of speakers, language inequalities manifest themselves as unequal access to information, communications, health care, security along with attenuated participation in political and civic life. These inequities are further exacerbated in a technological age, where only the most highly resourced (i.e. colonial) languages become the milieu for economic advancement (Odoje, 2013; Awobuluyi, 2016; Ganagana & Ogboru, 2019). Finally, there have been practical and technical challenges in language technology for indigenous languages like orthographic standardizations and consistent diacritic representation (Unicode) in electronic media and across device types.

For almost-extinct languages, machine translation offers hope 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). Using the new JW300 public dataset, we trained and evaluated baseline Neural Machine Translation (NMT) models for four of the widely spoken Edoid languages: Èdó, Èsán, Urhobo and Isoko.

1.1 EDOID LANGUAGES

Belonging to the eastern sub-branch of the Volta-Niger family within the Niger-Congo phylum, and spoken by approximately 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 Èdó, the most broadly spoken member language and the language of the famed Kingdom of Benin. Èdó, Èsán are members of the North-Central branch while Urhobo and Isoko belong to the South-Western family (Wikipedia, 2020). These languages were selected based on the availability of text and because they are the most widely spoken.

Edoid languages generally employ the SVO constituent order type, open syllable systems with very few consonant clusters. Each language has at least two basic tone levels, high (H) and low (L) with kinetic, downstepped or contour tones variously utilized. As tone patterns serve different lexical and grammatical functions, “the phonetic and phonological implementation of this system is in fact complex and difficult to pin down” (Rolle, 2013; Ogie, 2009; Adeniyi, 2010; Iloilo, 2013). Finally, nasalisation is very common for both vowels and consonants (Elugbe, 1989; Donwa-Ifode, 1986; Ikoyo-Eweto, 2018).

Within Nigeria there is scholarship on rule, phrase and statistical machine translation systems for majority tongues of Yorùbá, Igbo and Hausa (Odoje, 2013). The present study is the first work known to the authors done in computational linguistics for any of the Edoid languages, specifically for machine translation.

2 METHODOLOGY

We first built baseline models using the Transformer architecture, the dominant modeling approach for NMT. The Transformer uses an encoder-decoder structure with stacked multi-head self-attention and fully connected layers (Vaswani et al., 2017). Given the performance of Byte Pair Encoding (BPE) subword tokenization for low-resourced South African languages, we trained baseline models based the ablation study results by Martinus et al., some 4000 BPE tokens (Martinus & Abbott, 2019). Models were then re-trained for all four languages using the standard word-level tokenization.

2.1 DATASET

The recently published 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- $\{\text{Èdó, Èsán, Urhobo, Isoko}\}$ token pairs cardinality is itemized in Appendix Table 2. JW300 text is drawn from the Watchtower and Awake! religious magazines by Jehovah’s Witnesses (JW).

2.2 MODELS

The open-source, Python 3 machine translation toolkit `JoeyNMT` was used to train models based on the Transformer architecture (Kreutzer et al., 2019). Our training hardware was the commodity free-tier configuration on Google Colaboratory, a single core Xeon CPU instance and a Tesla K80 GPU. Model training elapsed over multiple days, as experiments were repeated for the different tokenization.

3 RESULTS

3.1 QUALITATIVE

Unsurprisingly, 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. Additional examples are listed in the Appendix.

3.2 ERROR ANALYSIS

While performing error analyses on the model predictions, we observed predictions that included dataset requires more preprocessing to remove scriptural chapter and verse text. chaptersverse names and figures. This will make the model more generally useful outside of religious text translations.

4 FUTURE WORK AND CONCLUSIONS

Fertile avenues for future work include investigating back-translation and different (subword) tokenization approaches as well as specific consideration of linguistic knowledge. We hope this initial effort will assist translators, bootstrap development and sustenance of scholarly and literary traditions and energize academic and industry interest in language technology for socio-linguistic and economic empowerment.¹ 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.

¹<https://github.com/Niger-Volta-LTI>

REFERENCES

- Harrison Adeniyi. Tone and nominalization in edo. *California Linguistic Notes*, 35(1):1–22, 2010.
- 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>.
- Oládélé Awobuluyi. Why We Should Develop Nigerian Languages. *Issues in Contemporary African Linguistics: A Festschrift for Oladele Awobuluyi*, 11:347, 2016.
- Shirley Donwa-Ifode. Phonetic variation in consonants (isoko). *Anthropological Linguistics*, 28(2):149–160, 1986. ISSN 00035483, 19446527. URL <http://www.jstor.org/stable/30028405>.
- Ben Ohiomamhe Elugbe. *Comparative Edoid: phonology and lexicon*. University of Port Harcourt Press, 1989.
- D. E. Peter Ganagana and Lawrence Efe Ogboru. Contrastive Study Of The Morphological Differences Between English, Izon And Isoko Languages. 2019.
- Evarista Ofure Ikoyo-Eweto. Phonetic differences between esan and selected edoid languages. *Journal Of Linguistics, Language and Culture*, 4(1), 2018.
- AKPOGHENE ONORIEVARIE Ilolo. *Vowel Reduction in Educated Isoko English*. PhD thesis, 2013.
- Julia Kreutzer, Joost Bastings, and Stefan Riezler. Joey NMT: A Minimalist NMT Toolkit for Novices. *To Appear in EMNLP-IJCNLP 2019: System Demonstrations*, Nov 2019.
- Laura Martinus and Jade Z. Abbott. A focus on neural machine translation for african languages. *CoRR*, abs/1906.05685, 2019. URL <http://arxiv.org/abs/1906.05685>.
- 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.
- Ota Ogie. *Multi-verb constructions in Edo*. Norges teknisk-naturvitenskapelige universitet, Det humanistiske fakultet , 2009.
- MO Omo-Ojugo. Esan language endangered. *Implications for the Teaching and Learning of Indigenous Languages in Nigeria*, 2004.
- Nicholas Rolle. Phonetics and phonology of Urhobo. *UC Berkeley PhonLab Annual Report*, 9(9), 2013.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.
- Wikipedia. Edoid languages — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Edoid_languages, 2020. [Online; accessed 07 February 2020].

A APPENDIX

Table 1: Per-language BLEU scores by BPE or word-level tokenization

Language	BPE		Word		Tokens	Sentences
	dev	test	dev	test		
Èdó	7.92	12.49	5.99	8.24	229,307	10,188
Èsán	4.94	6.25	3.39	5.30	87,025	4,128
Urhobo	15.91	28.82	11.80	22.39	519,981	214,546
Isoko	32.58	38.05	32.38	38.91	4,824,998	25,610

Table 2: Example Translations

Èdó	
Source:	Still , words of apology are a strong force toward making peace
Reference:	Ghele na , eme unu - uwou u re fi ob h gaga eva eruo udhedh
Prediction:	Ghele na , eme unu - uwou y gba ologbo n ma re ro ru udhedh
Source:	We can even ask God to create in us a pure heart
Reference:	Ma r sae tub yare ghn re k omai eva efuafo
Prediction:	Ma r sae tub yare ghn re ma omai eva efuafo
Èsán	
Source:	Still , words of apology are a strong force toward making peace
Reference:	Ghele na , eme unu - uwou u re fi ob h gaga eva eruo udhedh
Prediction:	Ghele na , eme unu - uwou y gba ologbo n ma re ro ru udhedh
Source:	We can even ask God to create in us a pure heart
Reference:	Ma r sae tub yare ghn re k omai eva efuafo
Prediction:	Ma r sae tub yare ghn re ma omai eva efuafo
Urhobo	
Source:	Still , words of apology are a strong force toward making peace
Reference:	Ghele na , eme unu - uwou u re fi ob h gaga eva eruo udhedh
Prediction:	Ghele na , eme unu - uwou y gba ologbo n ma re ro ru udhedh
Source:	We can even ask God to create in us a pure heart
Reference:	Ma r sae tub yare ghn re k omai eva efuafo
Prediction:	Ma r sae tub yare ghn re ma omai eva efuafo
Isoko	
Source:	Still , words of apology are a strong force toward making peace
Reference:	Ghele na , eme unu - uwou u re fi ob h gaga eva eruo udhedh
Prediction:	Ghele na , eme unu - uwou y gba ologbo n ma re ro ru udhedh
Source:	We can even ask God to create in us a pure heart
Reference:	Ma r sae tub yare ghn re k omai eva efuafo
Prediction:	Ma r sae tub yare ghn re ma omai eva efuafo