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, for centuries under colonial rule, Nigerian communities with primarily oral traditions have been the most susceptible to language endangerment (Rolle, 2013).

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). 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 translation models for four of the widely spoken Edoid languages: Ẹ̀dó, Ẹ̀sán, Urhobo and Isoko.

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. Ẹ̀dó, Ẹ̀sán are members of the North-Central branch while Urhobo and Isoko belong to the South-Western family (Wikipedia, 2020).

Edoid langauges are all tonal in nature, featuring multiple rising, falling and gliding tones for grammatical and lexical, syntacticpurposes toneach with a variety of dialects Esan is an analytic language with very little morphology, and maintains a canonical SVO word order.

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.

Kurdish is an Indo-European language mainly spoken in central and eastern Turkey, northern Iraq andSyria, and western Iran. It is a less-resourced language (Salavati and Ahmadi, 2018), in other words,a language for which general-purpose grammars and raw internet-based corpora are the main existingresources. The language is spoken in five main dialects, namely, Kurmanji (aka Northern Kurdish),Sorani (aka Central Kurdish), Southern Kurdish, Zazaki and Gorani (Haig and pengin, 2014).Creatinglexicaldatabasesandtextcorporaareessentialtasksinnaturallanguageprocessing(NLP)development. Text corpora are knowledge repositories which provide semantic descriptions of words. TheKurdish language lacks diverse corpora in both raw and annotated forms (Esmaili et al., 2013; Hassani,2018). According to the literature, there is no domain-specific corpus for Kurdish

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-{Èdó, Èsán, Urhobo, Isoko} token pairs cardinality is itemized in Table 1. {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

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 FUTURE WORK AND CONCLUSIONS

Avenues for future work include acquiring more data, hyper-parameter optimization and experiments with back-translation, tokenization approaches as well as folding in specific languistic information, say about the analytic nature of the morphology to help boost performance.

The objective of this paper was to report preliminary efforts to assist translators and the lay-person 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 broader language technology for socio-linguistic and economic empowerment.‘

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

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.

Nicholas Rolle. Phonetics and phonology of urhobo. *UC Berkeley PhonLab Annual Report*, 9(9), 2013.

Wikipedia. Edoid languages — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Edoid_languages, 2020. [Online; accessed 06-February-2020].

A APPENDIX

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

Language	**BPE**		**Word**		**Tokens**	
	dev	test	dev	test		
Èdó	7.92	12.49	5.99	8.24	229,307	10,188
Èṣán	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