TOWARDS NEURAL MACHINE TRANSLATION FOR EDOID LANGUAGES

Iroro Fred Ọ̀nọ̀mè Orife
Niger-Volta Language Technologies Institute
Seattle, Washington, USA
iroro@alumni.cmu.edu

Dr. John N. Orife
Indiana University of Pennsylvania,
Indiana, Pennsylvania, USA
jorife@iup.edu

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 the Edoid languages family of Southern Nigeria. Using public datasets, we trained and evaluated translation models for four widely spoken langauges in this group: Ẹ̀dó, Ẹ̀sán, Urhobo and Isoko. Trained models, code and datasets have been open-sourced to advance future research efforts on language technology for this minority language family.

1 INTRODUCTION

Language technology has an enabling effect on society. However computational linguistics research has only addressed about 1% of the world's languages. (Bird & Chiang, 2012). The disparity in

(Odoje, 2016)

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

Machine translation is relevant to the process of language documentation, preservation

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* comes from Ẹ̀dó, the language of the famed Kingdom of Benin and the most broadly spoken member. The two langauges under study, Urhobo and Isoko, are classified as South-Western Edoid, while Ẹ̀dó, Ẹ̀sán are classified as North-Central (Wikipedia).

Most African languages have been labelled "low resource". This is due in part to the limited body of academic research, lack of funding and online-datasets as well as low interest, due to preceptions about the prestige and utility of these languages in contemporary African life ODOJE; Awobuluyi (2016). Finally there are practical and technical challenges with respect to orthographic standardizations, consistent diacritic representation (Unicode) in electronic media and across device types.

work has been done on Edoid languages, this is in part due to the limited datasets, which is a function of a

The objective of this study is to carry out foundational work using current NMT techniques in language family with rich oral traditions but very little literature. We seek to energize interest in language technology research for these minority languages. 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 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. 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

Table 1: Evaluation BLEU scores

Language	**Word-level**		**BPE**		Training Tokens
	dev	test	dev	test	
Ẹ̀dó	10.0	0.01	30.1	11.8	100,000
Ẹ̀sán	11.22	20.4	30.9	11.9	300,300
Urhobo	11.33	1.342	33.4	11.2	3,000,000
Isoko	11.22	12.99	13.4	11.7	4,000,000

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 al-

gorithmic 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.

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.

ladele Awobuluyi. 26. why we should develop nigerian languages. *Issues in Contemporary African Linguistics: A Festschrift for Oladele Awobuluyi*, 11:347, 2016.

Steven Bird and David Chiang. Machine translation for language preservation. In *Proceedings of COLING 2012: Posters*, pp. 125–134, 2012.

CLEMENT ODOJE. Language inequality: Machine translation as the bridging bridge for african languages.

Clement Odoje. 12. the peculiar challenges of smt to african languages. *ICT, Globalisation and the Study of Languages and Linguistics in Africa*, pp. 223, 2016.

Wikipedia. Edoid languages. URL hhttps://en.wikipedia.org/wiki/Edoid_languages.

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