IGBO-ENGLISH MACHINE TRANSLATION: AN EVALUATION BENCHMARK

I. Ezeani, P. Rayson
Lancaster University,
Lancaster, UK

I. Onyenwe, C. Uchechukwu
Nnamdi Azikiwe University,
Awka, Nigeria

Mark Hepple
Sheffield University,
Sheffield, UK

1 INTRODUCTION

Although researchers are pushing the boundaries and enhancing the capacities of NLP tools and methods, works on African languages are lagging behind. A lot of focus on well-resourced languages such as English, Japanese, German, French, Russian, Mandarin Chinese etc. Over 97% of the world's 7000 languages, including African languages, are low-resourced for NLP i.e. they have little or no data, tools, and techniques for NLP research. For instance, only 5 out of 2965 (0.19%) authors of full-text papers in the ACL Anthology[1] extracted from the 5 major conferences in 2018 (ACL, NAACL, EMNLP, COLING and CoNLL) are affiliated to African institutions[2].

In this work, we discuss our effort toward building a standard evaluation benchmark dataset for Igbo-English machine translation tasks. Igbo[3] is one of the 3 major Nigerian languages spoken by over 50 million people globally, 50% of whom are in southeastern Nigeria. Igbo is low-resourced despite some efforts toward developing IgboNLP such as part-of-speech tagging: Onyenwe et al. (2014), Onyenwe et al. (2019); and diacritic restoration: Ezeani et al. (2016), Ezeani et al. (2018).

Although there are exiting sources for collecting Igbo monolingual and parallel data, such as the *OPUS Project* (Tiedemann (2012)) or the `JW.ORG`, they have certain limitations. The OPUS Project is a good source training data but, given that there are no human validations, may not be good as an evaluation benchmark. `JW.ORG` contents, on the other hand, are human generated and of good quality but the genre is often skewed to religious contexts and therefore may not be good for building a generalisable model.

This project focuses on creating and publicly releasing a standard evaluation benchmark dataset for Igbo-English machine translation research for the NLP research community. This project aims to build, maintain and publicly share a standard benchmark dataset for Igbo-English machine translation research . There are three key objectives:

1. Create a minimum of 10,000 English-Igbo human-level quality sentence pairs mostly from the news domain
2. To assemble and clean a minimum of 100,000 monolingual Igbo sentences, mostly from the news domain, as companion monolingual data for training MT models
3. To release the dataset to the research community as well as present it at a conference and publish a journal paper that details the processes involved.

2 METHODS

To achieve the objectives above, the task was broken down in the following phases:

Phase 1: Raw data collection and pre-processing:
This phase is to produce cleaned and pre-processed a minimum 10,000 sentences: 5,000 English and 5,000 Igbo. It involved the collection, cleaning and pre-processing (normalisation, diacritic restoration, spelling correction etc.) of Igbo and English sentences from freely available electronic

[1] `https://www.aclweb.org/anthology/`
[2] **Source:** `http://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/`
[3] **Igbo:** `https://en.wikipedia.org/wiki/Igbo_language`

texts (e.g. Wikipedia, CommonCrawl, local government materials, local TV/Radio stations etc).

Phase 2: Translation and correction
In this phase, the 10,000 sentence pairs are created manual translation and correction. The key tasks include:

1. Translating English sentences to Igbo (EN-IG)
2. Translating Igbo sentences to English (IG-EN)
3. Correcting the translations

5 Igbo speakers were engaged for the bidirectional of translations while 3 other Igbo speakers, including an Igbo linguist are assisting with the on-going corrections. Chunks ($\approx$ 250 each) of sentences are given to each translator in each direction (i.e. IG-EN and EN-IG). At the time of submission, we have $11,584$ sentence pairs as detailed in Table 1 while the splits of the parallel data into *development*, *text* and *hidden test* sets is shown in Table 2

Table 1: Breakdown of the Benchmark Evaluation Parallel Data

Type	**Sent pairs**	**Sources**
Igbo-English	5,836	`https://www.bbc.com/igbo`
English-Igbo	5,748	Mostly from local newspapers (e.g. Punch)
Total	$11,584$	

Table 2: Splits of the Benchmark Evaluation Parallel Data

Evaluation Splits	**IG-EN**	**EN-IG**
Development Set	5000	5000
Test set	500	500
Hidden Test	336	248

Phase 3: Manual checks and Inter-translator Agreement
This phase is currently on-going and it involves manually checking and correcting the 10,000 translated sentence pairs. This is to ensure that the translations conform with the contemporary communicative usage of the languages. Our approach so far is simplistic i.e. it seeks to establish absolute agreement between translators. We know it could overstate agreement (Lommel et al. (2014)), but we believe it will improve the quality of the translation. More work will be done in this area in future.

Phase 4: Monolingual Igbo sentence collection and pre-processing
The aim here is to collect and clean a minimum of 100,000 monolingual Igbo sentences. the cleaning process involves normalisation, diacritic restoration, spelling correction from freely available sources (news, government materials, Igbo literature, local TV/Radio stations etc).

A large chunk of the data is collected from the Jehova's Witness Igbo[4] contents. Though we included the Bible, more contemporary contents (books and magazine e.g. *Teta!* (*Awake!*), *Ulo Nche!* (*WatchTower*)) were the main focus. Also, we got contents from BBC-Igbo[5] and Igbo-Radio (`https://www.bbc.com/igbo`) as well as Igbo literary works(*Eze Goes To School*[6] and *Mmadu Ka A Na-Aria* by Chuma Okeke). This phase is still on-going but we have so far collected and cleaned $\approx$ 380k Igbo sentences as detailed in Table 3. It is important to point out that we have also collected data in other formats (e.g. audio, non-electronic texts) from local media houses which we hope to also transcribe and include in our collection.

[4]**Source:** `https://www.jw.org/ig/`
[5]`https://www.bbc.com/igbo/`
[6]`https://bit.ly/2vdGvKN`

Table 3: Data Sources and Counts

Source	Sentences	Tokens	UniqToks
eze-goes-to-school.txt	1272	25413	2616
mmadu-ka-a-na-aria.txt	2023	39731	3292
bbc-igbo.txt	34056	566804	28459
igbo-radio.txt	5131	191450	13391
jw-ot-igbo.txt	32251	712349	13417
jw-nt-igbo.txt	10334	253806	6731
jw-books.txt	142753	1879755	25617
jw-teta.txt	14097	196818	7689
jw-ulo-nche.txt	27760	392412	10868
jw-ulo-nche-naamu.txt	113772	1465663	17870
Total	**383,449**	**5,724,201**	**69,091**

3 ACCESS TO DATA

All data generated as described above are available under the Creative Commons license from this GitHub repository[7] and will be regularly updated.

4 CONCLUSION

This work presents an on-going project on building a benchmark evaluation dataset for Igbo–English machine translation project. The released dataset will hopefully be useful in fairly and more reliably comparing the performance of models built for IG-EN translations.

Our efforts in increasing the size of the sentence pairs as well as improving the quality of translations will continue in will be published as we progress. In addition to releasing the dataset to the research community, our plan for future works include building and comparing various machine translation models based on the current state-of-the-art methods. This will be followed by an in-depth analysis of their performances.

ACKNOWLEDGMENTS

The authors wish to acknowledge and thank Facebook AI Research (Facebook AI) for funding this project. Our immense gratitude also goes to Marc'Aurelio Ranzato and Francisco Guzmán for initiating, facilitating the funding and providing us with a lot of technical ideas.

REFERENCES

Ignatius Ezeani, Mark Hepple, and Ikechukwu Onyenwe. Automatic restoration of diacritics for igbo language. In *International Conference on Text, Speech, and Dialogue*, pp. 198–205. Springer, 2016.

Ignatius Ezeani, Ikechukwu Onyenwe, and Mark Hepple. Transferred embeddings for igbo similarity, analogy, and diacritic restoration tasks. In *Proceedings of the Third Workshop on Semantic Deep Learning*, pp. 30–38, 2018.

Arle Lommel, Maja Popovic, and Aljoscha Burchardt. Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*, 2014.

Ikechukwu Onyenwe, Chinedu Uchechukwu, and Mark Hepple. Part-of-speech tagset and corpus development for igbo, an african language. In *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, pp. 93–98, 2014.

[7] https://github.com/IgnatiusEzeani/IGBONLP/tree/master/ig_en_mt

Ikechukwu E Onyenwe, Mark Hepple, Uchechukwu Chinedu, and Ignatius Ezeani. Toward an effective igbo part-of-speech tagger. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):1–26, 2019.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.