IGBO-ENGLISH MACHINE TRANSLATION: AN EVALUATION BENCHMARK

Ignatius Ezeani[1], Ikechukwu Onyenwe[2], Chinedu Uchechukwu[2], Paul Rayson[1], Mark Hepple[3]

[1]Lancaster University UK, [2]Nnamdi Azikiwe University Nig, [3]The University of Sheffield UK

Introduction and Motivation

This is an on-going project with the IGBONLP initiative to contribute to developing AfricanNLP by building a standard evaluation benchmark dataset for Igbo-English machine translation tasks.
The motivation is drawn from the following:

- Only 0.19% of authors of full-text papers in 5 major conferences in the 2018 *ACL Anthology* are affiliated to African institutions[2]
- Therefore, African languages are typically low-resourced and lag behind in NLP research
- A major challenge is the lack of standard benchmark data for evaluating NLP systems
- Nigeria is home to ≈ 500 languages, a quarter of all African languages
- Igbo[1] language is a major Nigerian language with ≈ 50 million speakers globally

. Previous works on Igbo include:

- Part-of-speech tagger [5]
- Morphology Prediction [4]
- Diacritic Restoration [1]

Our Proposal

These sources provide IG-EN parallel data but with limitations:

- *OPUS Project* [6] is good source training data but with no human validations
- *JW.ORG* [3] is human-translated but in restricted (religious) genre

Aim: Build, Maintain and publicly share an Igbo-English Evaluation dataset for machine translation research Key objectives:

1. 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.

Also on-going but not concluded:

- **Building a baseline MT system - a key part of this proposal**
- To be presented in future work

Igbo Speaking Region of Nigeria

Fig. 1: Map of Nigeria Showing the Igbo Speaking Region.

Parallel Data Stats

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

Fig. 2: Tables 1 and 2 showing the Parallel Data and Split Stats

All data available here:
`https://github.com/IgnatiusEzeani/IGBONLP/tree/master/ig_en_mt`

Monolingual Data Stats

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**

Fig. 3: Table showing the Monolingual Data Stats

Acknowledgements

We acknowledge and thank Facebook AI Research (Facebook AI) for funding this project. We are also grateful to Marc'Aurelio Ranzato and Francisco Guzmán for continuously providing the technical support and ideas.

References

[1] 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*. 2018, pp. 30–38.

[2] *Geographic Diversity of NLP Conferences*. `http://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/`. Accessed: 13-04-2020.

[3] *JW.ORG: Ndiàmà Jèhóvà*. `https://www.jw.org/ig/`. Accessed: 13-04-2020.

[4] Ikechukwu E Onyenwe and Mark Hepple. "Predicting Morphologically-Complex Unknown Words in Igbo". In: *international conference on text, speech, and dialogue*. Springer. 2016, pp. 206–214.

[5] 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*. 2014, pp. 93–98.

[6] Jörg Tiedemann. "Parallel Data, Tools and Interfaces in OPUS". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012. ISBN: 978-2-9517408-7-7.