Machine-readable datasets for two low-resourced languages: Igbo & Nigerian Pidgin

Ebelechukwu Nwafor
Minh Phuc Nguyen
Department of Computer Science, Villanova University

1 Abstract

Currently, there is a lack of machine-readable datasets for two low-resource African languages: Igbo and Nigerian Pidgin. This obstacle hinders the progress of natural language processing (NLP) technologies in these languages despite their cultural and societal significance.
The primary goal of this research is to develop a machine-readable dataset for Igbo and Nigerian Pidgin. The dataset can be used to overcome the data gap and facilitate researchers and developers in building language models and other NLP tools in the future.

2 Background Info

Language translation, or machine translation (MT), is a fundamental sub-field of NLP that involves translating text or speech using computers.
Initially, two main methods for MT were tradition rule-based systems and statistical machine translation, but both were far from competent (Zhang & Zong, 2020).
Recently, neural machine translation (NMT) has revolutionized language translation as it utilizes deep learning techniques to model the translation process. (Tan et al., 2020)

With the advent of chatGPT-3, a large-scale language model, the boundaries of language translation have been pushed further. ChatGPT-3 has become a powerful tool for language translation. (Tan et al., 2020)

In this research project, the focus is on creating machine-readable datasets for low-resource African languages, Igbo and Nigerian Pidgin. The translation of these datasets to English using ChatGPT API showcases the potential of advanced NLP models to overcome data scarcity and support the development of language technologies for under-resourced languages.

3 Methodology

i. Data collection:
- Datasets were collected from 5 credible sources: BBC Igbo, BBC Pidgin, Ted Talk, Voice of Nigeria (national radio broadcaster), and Naijalingo (Nigerian Pidgin English dictionary).
- Webs were scraped using "BeautifulSoup" (Python library) and "concurrent. futures" module (concurrent programming).

ii. Data preprocessing:
- Techniques were used such as lines filtering, substrings removal, duplicate lines removal, and file I/O operations. After that, basic data analysis was performed.

iii. Data translation
- "text-davinci-003," an engine in chatGPT API, was used to translate the original dataset.
- Concurrent programming was used to speed up the process.

4 Results

In total, the two datasets (Igbo and Nigerian Pidgin) contain 238,150 sentences with an average of 19.4 words per sentence.

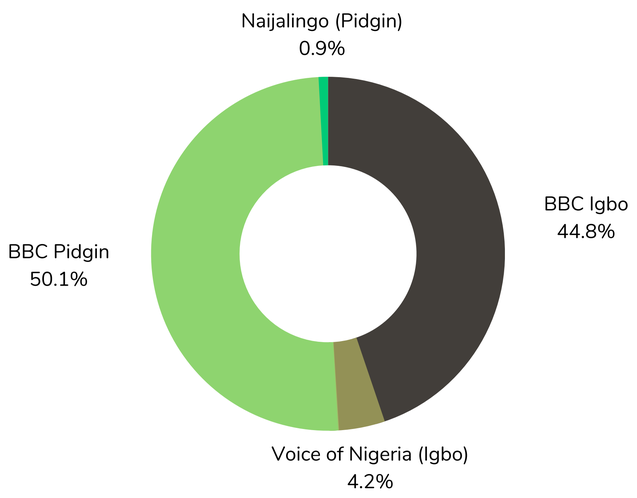
These are results after preprocessing the former datasets containing 2,000,000 sentences (collected from web scraping).

	# sentences after preprocessing	average words per sentence	# sentences translated by chatGPT-3
Nigerian Pidgin	121266	20.5	32428 (26.74%)
Igbo	116884	18.3	36993 (31.65%)
Total	238150	19.4	69421 (29.15%)

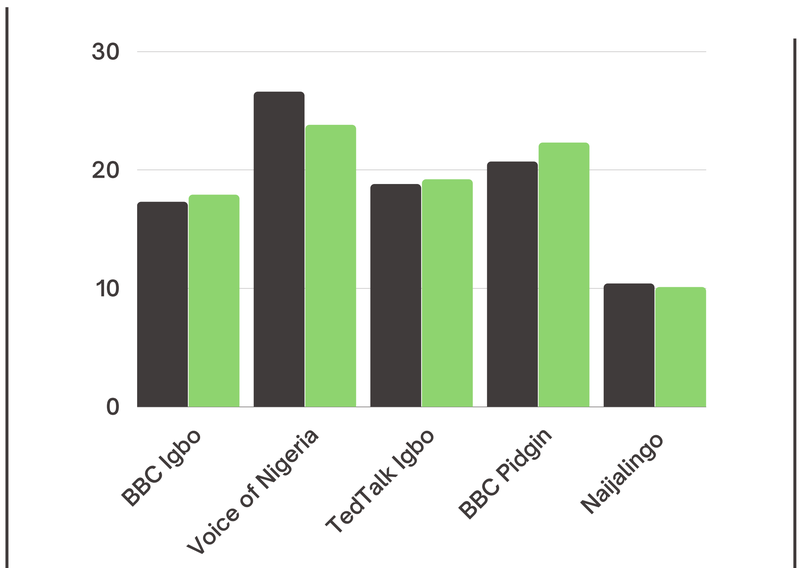
5 Analysis

With 121266 preprocessed Nigeria Pidgin sentences and 116884 preprocessed Igbo sentences, the research reaches the primary goal of creating machine-readable datasets for low-resource African languages.
The sources' distributions are highly disproportionate as BBC Pidgin and BBC Igbo consecutively takes about 50.1% and 44.8% of the total dataset.

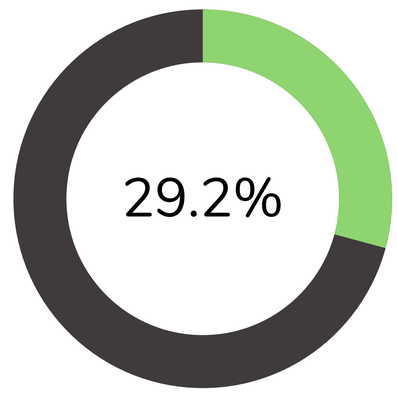
This can be explained by the lack of credible Igbo and Nigerian Pidgin webpages to extract text.
The average words per sentence in these datasets are relatively long (20.5 for Nigeria Pidgin sources and 18.3 for Igbo sources), demonstrating certain level of credibility of the datasets.
ChatGPT was able to generate translated sentences with about the same length as that of the original sentences. Only 29.2% of the dataset was translated by chatGPT due to the limited in quota of chatGPT API and billing.



The percentage of each source



Average words per sentence and per translated sentence for each source



Percentage of the total dataset being translated by chatGPT

6 Conclusion

The study helps address the lack of structured, machine-readable data for Igbo and Nigerian Pidgin by collecting, preprocessing, and translating data from credible sources. Furthermore, it suggests the use of advanced NLP models (ChatGPT) to overcome data scarcity. The work can facilitate future researchers in developing NLP tools for the two languages.

In the future, we plan to use Amazon's Mechanical Turk to label the dataset. We will then compare the translation efficiency between ChatGPT-3 model and non-expert annotators/translators to gain insights about chatGPT-3 model's translation quality.
Combining this insight with this study's results, we plan to build a translation model that would push the boundaries of language translation in the context of Igbo and Nigerian Pidgin.

Related literature
Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. AI Open, 1, 5–21. https://doi.org/10.1016/j.aiopen.2020.11.001 Z
Zhang, J., & Zong, C. (2020). Neural Machine Translation: challenges, progress and future. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2004.05809