Enhancing Inclusivity: Machine Translation of English Biological Research Papers to Pidgin-English (A language spoken by over 75 million people)

Flora Oladipupo, Anthony Soronnadi, Olubayo Adekanmbi
[Flora, Anthony, Olubayo]@datasciencenigeria.ai

ABSTRACT

The translation of scientific literature into local languages is vital for fostering an inclusive and widespread understanding of scientific knowledge. This ongoing research project focuses on developing and evaluating LLM's system designed to translate English biological research papers into Pidgin English. The work involves developing a corpus of biological datasets in Pidgin. Preliminary findings indicate that the system has the potential to significantly enhance the accessibility of scientific information for Pidgin-speaking communities, thereby promoting educational and research opportunities. Comparative evaluations using BLEU metrics is used to assess the effectiveness of state-of-the-art machine translation dataset specifically gotten from GPT 3.5, 4o-mini, 4.0 and 4o against our dataset.

INTRODUCTION

The translation of scientific literature is crucial for the dissemination of knowledge across linguistic and cultural boundaries. Pidgin, a Creole language spoken widely by 75 million people in West Africa Kasraee [2017], often serves as a bridge between different language speakers. However, scientific literature in Pidgin remains an underexplored area in AI research and academics. This study aims to develop a machine translation system to convert biological research papers written in English into Pidgin English, making scientific knowledge more accessible to Pidgin-speaking communities. Pidgin-English is the 4th most spoken language in Nigeria according to Statistica [2021].

The research goal is to evaluate LLMs Translation system on Biological dataset this is to bridge the science gap in Pidgin speaking communities. Such that a student whose native language is Pidgin-English in Delta, Nigeria have an adequate understanding of the same subject as a student in California, USA.

Figure 1: Image depiction showing how Knowledge and Comprehension can be established in a known language.

METHODOLOGY

The dataset used was open-source biological research papers. These papers are scraped and anonymized to protect authors' privacy. The papers was splitted into sentences giving a total of 845 rows of sentences which were then translated using the different GPT models and human translation. These are then evaluated using BLEU metrics.

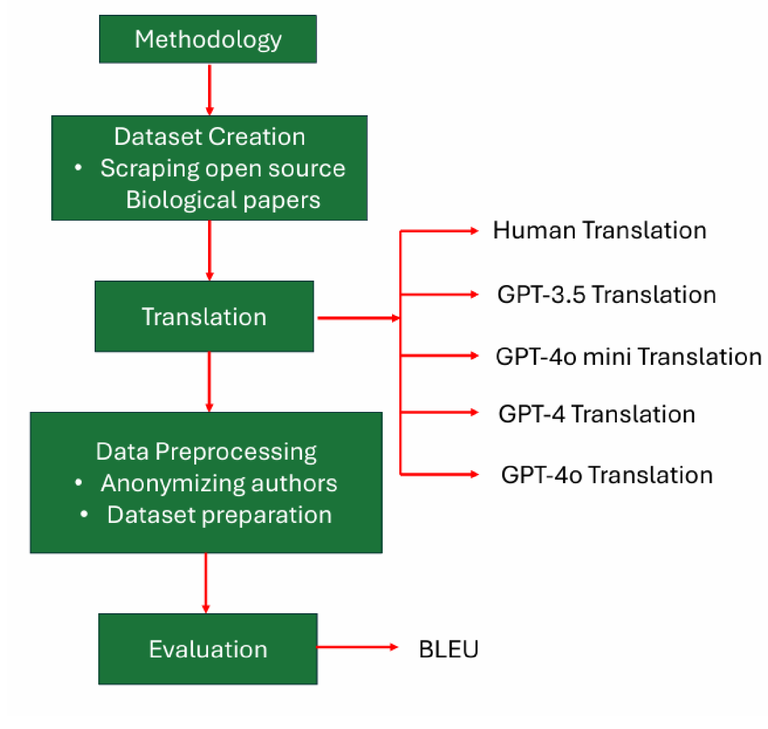

Figure 2: Methodology Flowchart

Table 1: Sample of Translated dataset

Text	Translation
English Text	CA can be conducted in vivo, in vitro, or ex vivo with any eukaryotic cell type, including vegetal tissue such as Allium cepa.
GPT-3.5 Translation	CA fit happen for inside body, for lab, or outside body with any cell wey get nucleus, including plants like Allium cepa.
GPT-4o mini Translation	CA fit be conducted in vivo, in vitro, or ex vivo with any eukaryotic cell type, including vegetal tissue like Allium cepa.
GPT-4.0 Translation	CA fit be conducted in vivo, in vitro, or ex vivo with any eukaryotic cell type, including vegetal tissue like Allium cepa.
GPT-4o Translation	CA fit conduct in vivo, in vitro, or ex vivo with any eukaryotic cell type, including vegetal tissue like Allium cepa.
Human Translation	CA fit dey happen for inside body, for lab, or outside body with any cell wey get nucleus (eukaryotic), it dey include plants like Allium cepa.

RESULT AND DISCUSSION

The evaluation of the system's performance in translating biological research papers was conducted by benchmarking the manual translation against various GPT-based large language models (LLMs), including GPT-3.5, GPT-4.0 mini, GPT-4.0, and GPT-4o. The performance metrics used in this assessment were BLEU, Loss, and Generated Length (Gen Len). The loss quantifies the difference between the model's translations and the actual target sentences. The BLEU score and loss metrics are critical indicators of the translation quality. As shown in Table 2, GPT-3.5 achieved the highest BLEU score and lowest Loss Score indicating that GPT-3.5 provides the most reliable and contextually accurate translations of biological papers.

Metric	GPT-3.5	GPT-4o mini	GPT-4.0	GPT-4o
BLEU	0.753079	0.241941	0.238126	0.344409
Loss	27.912944	58.859107	59.483391	50.684994

Table 2: BLEU and Loss Metrics score of the GPT models translation quality

Metric	Manual Translation	GPT-3.5	GPT-4o mini	GPT-4.0	GPT-4o
Gen Len	22.925544	22.427262	20.111111	20.099656	20.123711

Table 3: Generated Length of the GPT models translation against the Human Translation Length

CONCLUSION

This project has the potential to make scientific research more accessible across West African countries where Pidgin-English is spoken. Since the language one speaks predominantly plays a crucial role in their thought process therefore this research will aid in understanding biological science subjects better, thereby increasing research interest in biological academic research. This is an ongoing research in which the future work would include finetuning the dataset on multilingual machine translation models and applying the system to text-to-speech solutions, enabling Pidgin generated content in regions where this language is predominantly spoken.

REFERENCES

- Ebelechukwu Nwafor. A survey of machine translation tasks on Nigerian languages. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, page 6480–6486, 2022. URL https://aclanthology.org/2022.lrec-1.695.
- Zehila Babaci-Wilhite. The use of local languages for effective science literacy as a human right. In Human rights in language and STEM education, 2016.