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 crucial for promoting an inclusive and broad-based understanding of scientific knowledge. This research focuses on the evaluation of Eng-PidginBioData, a dataset specifically designed for translating biological research papers from English into Nigerian-Pidgin and development of the machine translation model. The study involves creating a comprehensive corpus of biological texts in Pidgin and finetuning of this dataset on machine translation models. Preliminary results suggest that this system has the potential to significantly improve the accessibility of scientific information for Pidgin-speaking communities, enhancing educational and research opportunities. The effectiveness of state-of-the-art machine translation models, including GPT-3.5, GPT-4.0-mini, and GPT-4.0, is evaluated using BLEU metrics, with comparative analysis performed against our dataset. The model finetuned on is the Pidgin UNMT and T5-model.

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].

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

RESULT AND DISCUSSION

The evaluation of an existing 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 and Loss. 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 1, 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.

The result of finetuning the dataset on the Pidgin UNMT model and T5 model gave an accuracy of 34.51 and 37.15 respectively with Pidgin UNMT a monolingual model doing better than T5 model a multilingual model. Training on more data and bigger batch size could improve the performance.

METHODOLOGY

The dataset used was open-source biological research papers. These papers are scraped and anonymized to protect authors' privacy. The papers were splitted into sentences giving a total of 2300 rows of sentences which were then translated using the different GPT models and human translation. These are then evaluated using BLEU metrics. The original dataset and the human translation equivalent were finetuned using the Pidgin UNMT model Ogueji and Ahia [2019].

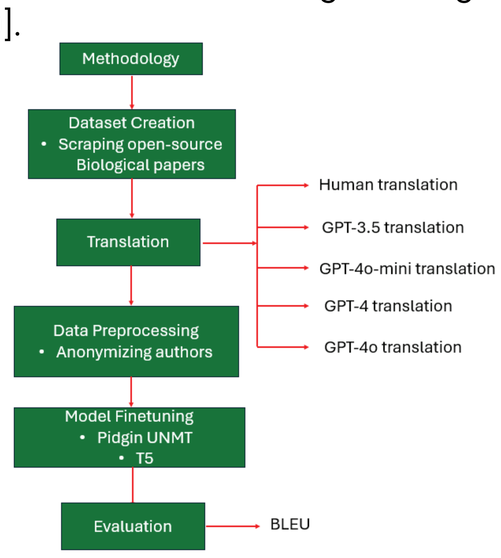

Figure 2: Methodology Flowchart

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.

Table 1: Sample of translated dataset

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

Model	Accuracy	Epoch	Pidgin-English BLEU	English-Pidgin BLEU
T5	34.51	20	37.72	35.02
Pidgin UNMT	37.15	39	60.75	33.00

Table 2: Metrics score of the GPT models translation quality

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 generating more data to improve the finetuning of the dataset on multilingual machine translation models and furthermore applying the system to text-to-speech solutions, enabling Pidgin generated content in regions where this language is predominantly spoken.