How Effective Are AI Models in Translating English Scientific Texts to Nigerian Pidgin: A Low-Resource Language?

Anonymous authors
Paper under double-blind review

Abstract

This research explores the challenges and limitations of applying deep learning models to the translation of scientific texts from English to Pidgin-English, a widely spoken but low-resource language in West Africa. Despite advancements in machine translation, translating domain-specific content such as biological research papers presents unique obstacles, including data scarcity, linguistic complexity, and model generalization issues. We investigate the performance of AI models, including Pidgin-UNMT, mt5-base model, AfriTeVa base, and Afri-mt5 base model through a comparative analysis using BLEU scores, CHRF, TER, Africomet metrics on a newly created Eng-PidginBioData dataset of biological texts. Our findings reveal significant gaps in model performance, emphasizing the need for more domain-specific fine-tuning, improved dataset creation, and collaboration with native speakers to enhance translation accuracy. By presenting real-world challenges encountered in applying deep learning to low-resource languages this research suggests strategies to overcome these barriers. Our study contributes valuable insights into how real-world constraints from limited data to domain mismatch continue to challenge the efficacy of AI-driven translation systems can be improved for underrepresented languages, offering actionable insights for more inclusive and effective scientific knowledge dissemination.

1 Introduction

Pidgin, a Creole language spoken widely by 75 million people in West Africa Kasraee (2017), often serves as a bridge between different language speakers. It is also the 4th most spoken language in Nigeria according to Statista (2021). Despite its widespread use, scientific literature in Pidgin remains underrepresented, limiting access to critical knowledge by speakers of this language. The translation of scientific literature is crucial for the dissemination of knowledge across linguistic and cultural boundaries. However, scientific literature in Pidgin remains an underexplored area in AI research and academics. This study focuses on development of a translation system for Pidgin language based on biological research papers, shedding light on how deep learning struggles to adapt to domain-specific and low-resource contexts.

Despite state-of-the-art performance in high-resource settings, AI-driven machine translation systems frequently fall short in low-resource scenarios. Data scarcity, contextual nuances, and scientific terminology complexity combine to limit the performance of existing Neural Machine Translation (NMT) models. In this paper, we explore machine translation in the domain of biological research texts. Our contributions are as follows:

- **Dataset Creation**: We present **Eng-PidginBioData**, a domain-specific high quality English-to-Pidgin parallel corpus focusing on biological research texts.
- **Comparative Analysis**: We evaluate **mt5 base**, **AfriTeVa base**, **Afri-mt5 base** and **Pidgin UNMT** models using BLEU scores metrics, highlighting performance gaps and unexpected failures.
- **Insights & Strategies**: We identify real-world challenges (limited data, cultural nuances, scientific jargon) and propose practical interventions (domain adaptation, expanded corpora, native speaker collaboration) to narrow the performance gap.

2 Related Works

Using local languages in science education improves comprehension, engagement, and equity Babaci-Wilhite (2016). Despite advances in Neural Machine Translation (NMT), translation models still struggle in low-resource settings like Nigerian Pidgin (pcm) due to the lack of large-scale parallel corpora Nwafor (2022). A breakthrough came with Pidgin-UNMT Ogueji & Ahia (2019), which used unsupervised learning techniques to train an NMT system without parallel data, paving the way for Creole language translation research. The Afri-mt5 model Adelani et al. (2022), an adaptation of mT5 for African languages, showed limited performance on Pidgin due to its French-heavy training data. Studies suggest fine-tuning with in-domain data improves translation quality, yet Afri-mt5 remains constrained in technical domains. AfriTeVa Oladipo et al. (2023) advanced African NLP but underperformed in Pidgin-English MT due to its lack of a dedicated Pidgin corpus. Cheetah Adebara et al. (2024), supporting 517 African languages, achieved a BLEU score of 32.64 for English-to-Pidgin MT but lacked domain-specific optimization. Toucan Elmadany et al. (2024), an enhancement of Cheetah, was fine-tuned on AfroLingu-MT, Africa's largest MT benchmark. While effective for general Pidgin-English translation, it still lacks fine-tuning for specialized domains like biomedical, technical, and legal texts.

3 Methodology

3.1 Dataset

The Nigerian Pidgin-English dataset used in this study was meticulously curated by scraping open-sourced English biological research papers. To maintain privacy, the authors' names were removed from the scraped data through an anonymization process. This corpus was then segmented into 2,300 sentences, forming the foundational dataset for translation. These sentences were then translated manually into Pidgin-English. To ensure linguistic accuracy and cultural relevance, these translations underwent a manual correction process by human annotators proficient in Nigerian Pidgin. This dual-step translation process aimed to refine the quality of the dataset, making it a reliable benchmark for evaluating machine translation models. This dataset is called **Eng-PidginBioData**.

Split	Size	TTR (English)	TTR (Pidgin)
Train	1380	16.6378	11.9739
Dev	460	27.5226	19.1845
Test	460	28.4705	20.0848

Table 1: Type-Token Ratio (TTR) for English and Pidgin on Eng-PidginBioData

To assess the linguistic richness and diversity of the curated dataset Eng-PidginBioData, we computed the Type-Token Ratio (TTR) for both English and Pidgin-English text across the training, development, and test splits. The results show that TTR values are consistently lower in Pidgin-English compared to English, reflecting Pidgin's characteristic reliance on a smaller, more flexible vocabulary with frequent code-switching and word reuse. Specifically, in the training set, TTR for English is 16.64, whereas Pidgin-English has a lower TTR of 11.97, suggesting that Pidgin exhibits more lexical repetition. Similarly, in the development and test sets, TTR values for English (27.52 and 28.47) remain higher than for Pidgin (19.18 and 20.08), reinforcing this pattern.

3.2 Model Fine-Tuning

The Eng-PidginBioData dataset was used to fine-tune existing machine translation models. The models includes: **Pidgin-UNMT 220M parameters**, Ogueji & Ahia (2019), **mt5 base 580M parameters** Raffel et al. (2023), **AfriTeVa base 229M parameters** Oladipo et al. (2023), **Afri-mt5 base 580M parameters** Adelani et al. (2022)

A uniform set of hyperparameters was used to ensure consistency across all models. Specifically, a learning rate of 2e-5 was employed with AdamW optimization, and a batch size of 4 was used. Additionally, a maximum sequence length of 256 tokens was maintained throughout training. These hyperparameter choices align with findings from Nag et al. (2024), who demonstrated the effectiveness of low learning rates in fine-tuning transformer models for low-resource languages.

The models were evaluated using BLEU metrics, providing insights into their translation accuracy and linguistic fluency.

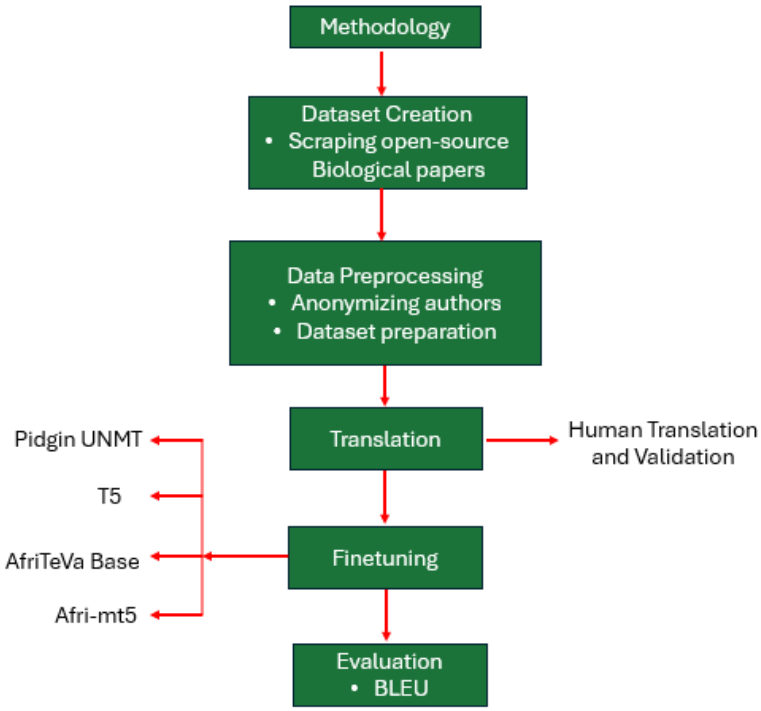

Figure 1: Methodology Flowchart

4 Result and Discussion

4.1 Analysis

Model	BLEU	TER	CHRF	AfriComet
Afri-mt5-base	23.37	66.09	43.46	0.42
AfriTeVa-base	25.55	53.03	59.05	0.51
Pidgin-UNMT-base	30.34	35.02	64.59	0.71
mt5-base	0.03	95.54	8.33	0.22

Table 2: Performance metrics for different models on Eng-PidginBioData Dataset

The table presents the evaluation of various machine translation models on the Eng-PidginBioData dataset, incorporating BLEU, Translation Edit Rate (TER), CHRF, and AfriComet metrics to assess their translation accuracy, fluency, and semantic preservation. The results highlight significant disparities in performance, particularly between the base mT5 model and AfriMT5, an adaptation built specifically for African languages. The results reveal that mT5-base performed exceptionally poorly across all metrics, achieving a BLEU score of just 0.03 and the highest TER (95.54), indicating nearly complete misalignment with reference translations. The CHRF score (8.33) and AfriComet score (0.22) further confirm its inability to generate meaningful Pidgin-English translations. Despite mT5 being a multilingual model trained on a diverse set of languages, its performance on Pidgin-English is drastically low, suggesting that Pidgin-English was either underrepresented or absent in its pretraining corpus. This highlights a major limitation of general multilingual models when applied to low-resource languages that lack adequate pretraining data. AfriMT5, which was built upon mT5 and adapted for African languages, significantly outperformed its base model,

achieving a BLEU score of 23.37 compared to mT5's near-zero score. This dramatic improvement suggests that fine-tuning mT5 with African language data is essential for improving its performance on low-resource languages such as Pidgin-English. However, AfriMT5 still struggled in comparison to Pidgin-UNMT and AfriTeVa, exhibiting a high TER (66.09) and low CHRF (43.46), indicating limited fluency and high translation errors. This suggests that despite adaptation, AfriMT5 does not yet fully capture the nuances of Pidgin-English, especially for scientific translations. Among the tested models, Pidgin-UNMT achieved the highest BLEU (30.34), lowest TER (35.02), and the best AfriComet score (0.71), confirming that Pidgin-specific models remain the most effective for this translation task. AfriTeVa, although slightly behind Pidgin-UNMT, demonstrated better fluency than AfriMT5 (CHRF: 59.05 vs. 43.46) and lower translation errors (TER: 53.03 vs. 66.09). This suggests that while AfriTeVa is not optimized for Pidgin-English, its pretraining on African languages provided some advantages over AfriMT5.

4.2 Challenges and Insights

One of the major challenges encountered in this study was the lack of large-scale, high-quality parallel corpora for Pidgin-English, particularly in the biological domains. The absence of well-structured datasets made it necessary to curate domain-specific data manually, a process that involved scraping, translating, and human-correcting scientific texts. This data limitation directly impacted the fine-tuning efficiency of the models, as they lacked extensive Pidgin-English text exposure during pretraining. A key observation from this study is the significant disparity in performance between Pidgin-specific models and adapted multilingual models. AfriMT5, despite being an African language adaptation of mT5, struggled with fluency and accuracy, achieving a high TER (66.09) and lower CHRF (43.46) than AfriTeVa. These results suggest that African multilingual models require additional domain adaptation to effectively translate scientific texts into Pidgin-English. Another observation during evaluation was that Pidgin-English translations frequently mixed English words where no direct Pidgin equivalent existed. While this is common in natural Pidgin usage, it posed a challenge for evaluation metrics such as BLEU, which struggle with code-mixed sentences. The lack of a standardized Pidgin-English orthography further complicated model evaluation, as multiple valid translations could exist for the same phrase, leading to inconsistencies in scoring. Future models should incorporate context-aware embeddings and specialized tokenization strategies to handle code-switching more effectively. Also, while automatic metrics provided insights into model performance, the study lacked real-time human feedback loops to assess fluency, coherence, and cultural relevance. Incorporating human evaluators for post-editing and ranking translations could further refine the models. To enhance translation quality, future research should focus on expanding the dataset with domain-specific glossaries, fine-tuning models with a larger Pidgin-English scientific corpus, and incorporating transfer learning techniques. Also, introducing human-in-the-loop translation validation for refining Pidgin-English translations in real-world applications could further improve scientific text translation accuracy, ensuring that Pidgin speakers can access critical information in education, and research.

4.3 Conclusion

This study evaluated fine-tuned machine translation models for scientific Pidgin-English, highlighting the strengths and limitations of multilingual and African language-adapted models. Pidgin-UNMT emerged as the most effective for Pidgin-to-English (BLEU: 30.34, AfriComet: 0.71), while mT5 performed poorly (BLEU: 0.03), emphasizing the need for targeted fine-tuning on low-resource languages. While AfriMT5 and AfriTeVa showed improvements through adaptation, they struggled with fluency, semantic adequacy, and technical terms. High TER scores and evaluation challenges from code-switching and non-standardized orthography further highlight the limitations of current automatic metrics. Future research should explore hybrid models, context-aware embeddings, and expanded datasets to enhance scientific Pidgin-English translation. Addressing these challenges will improve machine translation for scientific literacy, education, and accessibility for Pidgin-English speakers.

REFERENCES

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. Cheetah: Natural language generation for 517 African languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12798–12823, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.691.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223.

Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith. Massively multilingual word embeddings. *arXiv preprint*, 2016.

Artetxe, G. Labaka, and E. Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294, 2016.

Artetxe, G. Labaka, and E. Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462, 2017a.

M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation. *arXiv preprint*, 2017b.

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.

Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Conneau, G. Lample, M. A. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *arXiv preprint*, 2017.

Abdelrahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. Toucan: Many-to-many translation for 150 african language pairs. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 13189–13206, 2024.

Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. *arXiv preprint*, 2018.

Najiba Kasraee. Bbc blogs - academy - working towards a standard pidgin, 2017. URL https://www.bbc.co.uk/blogs/academy/entries/70f2a30c-40a5-463c-9480-1d63e7d5f44a. Accessed: 2024-05-21.

G. Lample, A. Conneau, L. Denoyer, and M. A. Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint*, 2017.

G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. A. Ranzato. Phrase-based & neural unsupervised machine translation. *arXiv preprint*, 2018.

Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint*, 2013a.

Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013b.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. Efficient continual pre-training of llms for low-resource languages, 2024. URL https://arxiv.org/abs/2412.10244.

Ebelechukwu Nwafor. A survey of machine translation tasks on nigerian languages. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6480–6486, 2022. URL https://aclanthology.org/2022.lrec-1.695.

Kelechi Ogueji and Orevaoghene Ahia. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. 2019.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. Better quality pre-training data and t5 models for African languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 158–168, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.11.

Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Peter and H. G. Wolf. A comparison of the varieties of west african pidgin english. *World Englishes*, 26(1):3–21, 2007.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Ravi and K. Knight. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 12–21, 2011.

Ruder, I. Vulić, and A. Søgaard. A survey of cross-lingual word embedding models. *arXiv preprint*, 2017.

S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint*, 2017.

Statista. Population in nigeria by languages spoken, 2021. URL https://www.statista.com/statistics/1285383/population-in-nigeria-by-languages-spoken/. Accessed: Feb 4, 2025.