BRIDGING THE LANGUAGE GAP: EVALUATING MACHINE TRANSLATION FOR ANIMAL HEALTH IN LOW-RESOURCE SETTINGS

Godwin Adegbehingbe, Anthony Soronnadi, Ife Adebara ,Olubayo Adekanmbi
Research & Innovation Department
Data Science Nigeria
Lagos, Nigeria
{Godwin,Anthony,Ife,Olubayo}@datasciencenigeria.ai

ABSTRACT

Machine translation (MT) has made significant progress in high-resource languages, but translating technical texts into low-resource languages remains an open challenge. This study investigates the ability of state-of-the-art multilingual models to translate animal health reports from English to Yorùbá, a crucial task for enhancing veterinary communication in underserved regions. Although previous research has explored low-resource MT, domain-specific translation for animal health has been largely overlooked. Using a curated dataset of 1,468 parallel sentences, we evaluated several MT models in zero-shot and fine-tuned settings. Despite the promise of multilingual models, we find substantial limitations in their ability to generalize to this domain, raising concerns about their applicability in specialized, low-resource contexts. We analyze potential causes, including vocabulary mismatch, training data scarcity, and constraints of model architecture. Our findings highlight the need for more targeted approaches to low-resource domain-specific MT and emphasize the broader implications for AI deployment in real-world applications.

1 INTRODUCTION

Machine translation (MT) has transformed global communication, yet its promise of equitable access remains largely unfulfilled for low-resource languages like Yorùbá. Specialized domains,such as animal health reporting, still face significant neglect despite their critical importance(Bisoye, 2020).In Yorùbá-speaking regions of Nigeria, livestock farming sustains over 40% of rural livelihoods (Idowu & Babalobi, 2020) ,and timely translation of animal health reports is essential for monitoring zoonotic disease outbreaks, safeguarding food security, and ensuring effective veterinary care (Food and Agriculture Organization, 1999). Despite the crucial role of animal health reporting in these communities, existing MT systems struggle to convey veterinary terminology accurately(Ranathunga et al., 2022). Generic models such as No Language Left Behind (NLLB), ChatGPT, and Google Translate frequently produce mistranslations that undermine their utility in real-world scenarios (Nwafor & Andy, 2022). For example, translating key veterinary terms such as 'Avian influenza' into Yorùbá frequently results in semantic inaccuracies. Systems such as ChatGPT incorrectly render it as "ijàpá apáta" (rock tortoise), while NLLB generates "àrùn àrífín eye" (meaning "bird disrespect disease")—errors that obscure the disease's virological nature. These errors are further compounded by Yorùbá's tonal orthography,where slight variations in pitch drastically alter meaning (Abenet, 2024). Such failures pose serious risks, including misdiagnoses and delayed containment of outbreaks like Rabies, where mistranslations as "central nervous system" or misidentifications of animal species (e.g., "fox" rendered as "kìnnìún" meaning lion) could derail public health responses.

The root of this challenge lies in the mismatch between generic MT training data and the requirements of specialized domains. While transfer learning and domain adaptation have improved domain-specific MT in high-resource languages (Zoph et al., 2016), their effectiveness for Yorùbá remains largely untested due to limited available corpora and a lack of systematic evaluation in

specialized fields like veterinary medicine (Adelani et al., 2021). Prior work on low-resource MT has primarily focused on general-purpose translation (Adelani et al., 2021), leaving domain-specific applications underexplored(Dew et al., 2018). MT systems, predominantly trained on broad, non-specialized corpora, often struggle with the specialized vocabulary and contextual nuance required for veterinary texts. To address these challenges, this study introduces VetYorùbá, a high-quality, expert-validated parallel dataset designed to evaluate MT models in animal health reporting. Using this dataset, we assess state-of-the-art multilingual models under zero-shot and fine-tuned conditions, revealing that fine-tuning significantly improves translation quality but fails to resolve errors in critical terminology fully.Unlike general studies on low-resource MT, this work underscores the importance of domain adaptation beyond conventional fine-tuning, advocating for novel approaches that integrate linguistic features unique to Yorùbá. Our findings have broader implications for deploying MT in real-world, high-stakes applications, emphasizing the need for customized architectures that accommodate the complexities of African languages.

2 Related Work

Recent advances in machine translation (MT) have significantly improved low-resource language translation through transfer learning and unsupervised MT techniques. For African languages, particularly Yorùbá, pre-trained multilingual models like mT5 and mBART have shown promising results when fine-tuned on Yorùbá data (Adelani et al., 2022). However, challenges persist in domain-specific applications, especially in specialized fields such as animal health, where standardized terminologies are often absent or underdeveloped (**?**). Existing MT systems like NLLB and Google Translate frequently produce erroneous translations of technical terms, highlighting the need for domain-specific fine-tuning (Adebara & Abdul-Mageed, 2022).To address data scarcity in low-resource MT systems, researchers have explored various augmentation techniques. Back-translation has shown promise by creating synthetic parallel data from monolingual target-language content(Jauregi Unanue & Piccardi, 2020), though its effectiveness in preserving technical accuracy remains uncertain for domain-specific translations(Baruah & Singh, 2022). Additionally, synthetic data generation techniques have been investigated for neural MT (Tonja et al., 2023), while human-in-the-loop strategies incorporating domain experts (Nunes Vieira, 2019) have emerged as crucial approaches for improving translation quality, particularly in specialized domains (Yang et al., 2023).Evaluation of MT systems in specialized domains requires comprehensive assessment approaches that go beyond traditional metrics. While metrics such as BLEU, AfriComet and chrF provide insights into different aspects of translation quality, (Zappatore & Ruggieri, 2023) argue that specialized domains like biomedical MT require tailored evaluation strategies emphasizing terminology accuracy and practical usability. For Yorùbá animal health translation, these metrics collectively offer a multi-faceted assessment framework: BLEU measures n-gram overlap, AfriComet accounts for semantic accuracy in African languages, and chrF captures character-level precision, particularly valuable for morphologically rich languages like Yorùbá.

3 Dataset

We introduce VetYorùbá, a specialized parallel corpus for animal health reporting consisting of 1,468 manually verified English-Yorùbá sentence pairs. The corpus was strategically curated to capture the linguistic and terminological complexity inherent in veterinary documentation, focusing on disease surveillance and outbreak reporting in South-West Nigeria.

Split	Size	TTR (English)	TTR (Yoruba)
Train	1172	0.2243	0.1672
Dev	147	0.4706	0.3629
Test	147	0.4592	0.3485

Table 1: Dataset split and Type-Token Ratio (TTR) for English and Yoruba sentences

We collected our data from three primary sources: the World Organisation for Animal Health (WOAH) reports focusing on seven epidemiologically significant diseases in the region: Rabies, Avian Influenza, Newcastle Disease, Foot-and-Mouth Disease (FMD), African Swine Fever (ASF),

Bovine Tuberculosis, and Peste des Petits Ruminants (PPR). Food and Agriculture Organization (FAO) documentation covering animal health practices, preventive measures, and outbreak management protocols, selected to enhance the corpus's terminological breadth. Real-time epidemiological data extracted using PADI-Web (Valentin et al., 2020), an event-based surveillance tool that aggregates information from both structured (official reports) and unstructured sources (news articles, social media).

We focused on maintaining a balanced representation across different disease contexts and livestock categories.We got native speakers of Yorùbá to translate the sentences. To address limitations like inconsistent rendering of technical terminology, structural anomalies in complex veterinary descriptions, omission or misinterpretation of specialized medical terms, we implemented a comprehensive post-editing protocol, a veterinary doctor fluent in both languages reviewed and corrected the translations, focusing on accurate translation of technical terminology. This human-in-the-loop approach proved crucial for maintaining both domain expertise and linguistic accuracy (Yang et al., 2023), particularly in cases where direct translation of veterinary terminology required careful consideration of cultural and linguistic nuances in Yorùbá.For data partitioning, we used an 80–10–10 split by first dividing the dataset into training and test sets (80% and 20%, respectively) using a random state of 42, and subsequently splitting the 20% equally into validation and test sets.

4 EXPERIMENTAL SETUP

The experimental setup for our study was designed to evaluate the performance of various machine translation models on a domain-specific task involving Yorùbá animal health reports. We began with a curated dataset of 1,468 parallel English–Yorùbá sentences. The raw data underwent standard preprocessing procedures, including cleaning, tokenization, and normalization, to ensure consistency and to handle the specific challenges posed by Yorùbá's tonal orthography. We conducted experiments in both zero-shot and fine-tuning settings. In the zero-shot experiments, the pre-trained models were evaluated directly on the test set without any further adaptation to the domain-specific data. In our fine-tuning phase, we employed a uniform set of hyperparameters across all models: a maximum sequence length of 256 tokens, a batch size of 4, a learning rate of 2e-5, and training for 10 epochs. These choices were informed by the findings of (Verma & Kolhatkar, 2023), who conducted a hyperparameter ablation study to optimize transformer-based machine translation models for single GPU training. Their research identified optimal hyperparameter configurations that balance model performance and computational efficiency, guiding our selection to ensure effective training on a T4 GPU. The experiments were implemented using PyTorch and the HuggingFace Transformers library, ensuring both efficiency and reproducibility. The performance of each model was measured using a combination of automatic evaluation metrics, including BLEU(Papineni et al., 2002), chrF(Popović, 2015), and AfriComet(Wang et al., 2024), to capture various dimensions of translation quality, such as n-gram overlap, character-level precision.Human evaluation was also conducted to qualitatively assess translation fluency and adequacy. This comprehensive experimental setup allowed us to robustly compare the zero-shot capabilities and the benefits of fine-tuning for machine translation in a specialized, low-resource domain.For our experiments, we evaluated the following models: AfriMT5_base (Adelani et al., 2022), AfriTeVa_base Jude Ogundepo et al. (2022), AfriTeVa_v2_base Oladipo et al. (2023), mT0base (Muennighoff et al., 2023), mT5_base (Xue et al., 2021), and NLLB 3.3B (NLLB Team et al., 2022).

5 RESULT

The performance of the evaluated machine translation models was quantified using BLEU (Papineni et al., 2002), chrF (Popović, 2015), and AfriComet (Wang et al., 2024) metrics under both zero-shot and fine-tuned conditions. Overall, fine-tuning on our domain-specific dataset of 1,468 English–Yorùbá sentence pairs resulted in marked improvements across all metrics. In the zero-shot setting, the models generally exhibited low performance, with many struggling to produce coherent translations in the specialized domain of animal health. mT0 achieved a BLEU score of 11.57, while other models such aAfriMT5_base and AfriTeVa_v2 recorded near-zero BLEU scores (0.0003 and 0.005, respectively). However, the NLLB model demonstrated relatively better zero-shot performance with a BLEU score of 2.9 and comparable scores in chrF, indicating some capacity for handling the domain despite the lack of fine-tuning. Fine-tuning of the models on the curated

Model	ZS BLEU	FT BLEU	ZS chrF	FT chrF	ZS AfriComet	FT AfriComet
AfriTeVa_base	0.086	**13.2**	3.634	**37.49**	15.8	35.0
AfriTeVa_v2	0.005	0.03	1.226	3.8	16	13.0
Afri-mT5	0.0003	0.9	2.792	17.65	11.39	18.0
mT0	11.57	**15.9**	**32.06**	**44.86**	**54.9**	47.0
mT5	0.071	0.08	8.137	6.48	22.1	10.0
NLLB 3.3B	2.9	**45.89**	19.47	**66.85**	44.3	**62.0**

Table 2: Translation performance of models in zero-shot (ZS) and fine-tuned (FT) settings. Higher values indicate better translation quality across all three metrics: BLEU measures n-gram overlap, chrF captures character-level precision, and AfriComet evaluates semantic alignment."

veterinary dataset significantly improved translation quality. The BLEU score of the mT0 model improved to 15.9, while NLLB 3.3B exhibited the most dramatic gain, rising from 2.9 to 45.89. This improvement was consistently reflected in the chrF scores, with NLLB 3.3B increasing from 19.47 to 66.85. The AfriComet metric further supported these improvements, particularly for the NLLB 3.3B and the AfriTeVa base, whose fine-tuned scores of 62 and 35, respectively, signified better semantic alignment and contextual accuracy in translations. These results underscore the importance of domain-specific fine-tuning to enhance the performance of MT systems in low-resource and specialized settings. The substantial improvements observed in key models, particularly NLLB 3.3B, confirm that fine-tuning can mitigate the limitations of zero-shot translation and lead to more accurate and reliable translations of technical content in Yorùbá.

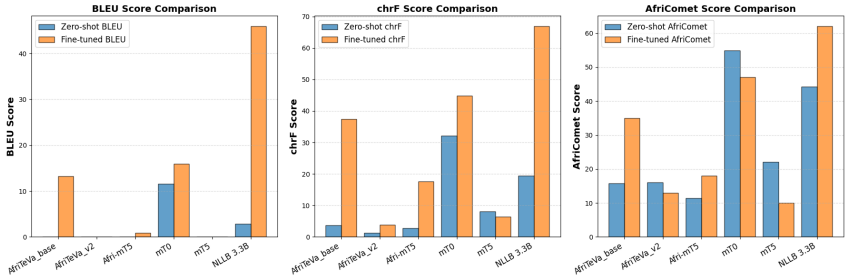

Figure 1: MT Models Performance on Yorùbá Animal Health Translation

6 Conclusion

This study explored the challenges of applying machine translation (MT) models to the domain-specific task of translating animal health reports into Yorùbá. It also introduces VetYorùbá, a high-quality, expert-validated parallel dataset for Yoruba animal health translation. Through systematic evaluation, we demonstrated that fine-tuning significantly improves translation performance, reinforcing the importance of domain adaptation for specialized MT tasks.Our results reveal substantial improvements across all evaluated models after fine-tuning our dataset of English-Yorùbá sentence pairs. The NLLB 3.3B model exhibited the highest performance gains. Other models, such as mT0, also demonstrated meaningful improvements, while models that initially produced near-zero scores (AfrimT5, AfriTeVa_v2) showed moderate but notable progress after fine-tuning. These improvements were consistently reflected across all evaluation metrics.Despite these gains, our findings reveal persistent gaps in domain-specific terminology adaptation, which emphasizes the limitations of fine-tuning alone in handling complex linguistic features such as tonal variations and specialized veterinary terminology. Given the limited dataset size, we do not claim definitive conclusions on the overall effectiveness of fine-tuning but rather highlight the challenges that remain for Yoruba MT in the animal health domain.Our work provides a benchmark dataset and evaluation framework that future studies can build upon. Expanding VetYorùbá with more parallel data and refining human-in-the-loop validation techniques are key next steps toward improving translation quality. This research highlights the need for Machine translation approaches that integrate linguistic and domain-specific knowledge, moving beyond conventional fine-tuning to better support underserved linguistic communities in critical fields such as public health and veterinary medicine.

REFERENCES

Tadesse Alemu Abenet. Bridging the gap: Legal and medical translation in african indigenous languages. *European Journal of Multilingualism and Translation Studies*, 4(1), 2024. doi: 10.46827/ejmts.v4i1.497. URL https://oapub.org/lit/index.php/EJMTS/article/view/497.

I. Adebara and M. Abdul-Mageed. Towards afrocentric nlp for african languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3814–3841, Dublin, Ireland, 2022. Association for Computational Linguistics.

D. I. Adelani, J. O. Alabi, A. Fan, J. Kreutzer, X. Shen, M. Reid, D. Ruiter, D. Klakow, and P. Nabende. The effect of domain and diacritics in Yoruba-English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pp. 61–75, Virtual, 2021. Association for Machine Translation in the Americas.

D. I. Adelani, J. O. Alabi, A. Fan, J. Kreutzer, X. Shen, M. Reid, D. Ruiter, D. Klakow, P. Nabende, E. Chang, et al. A few thousand translations go a long way! Leveraging pre-trained models for African news translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2022.

Rupjyoti Baruah and Anil Kumar Singh. *A Clinical Practice by Machine Translation on Low Resource Languages*. CRC Press eBooks, 2022.

E. Bisoye. Machine translation and Yorùbá: Matters arising. *Unilag Journal of Humanities*, 8(2), 2020.

K. N. Dew, A. M. Turner, Y. K. Choi, A. Bosold, and K. Kirchhoff. Development of machine translation technology for assisting health communication: A systematic review. *Journal of Biomedical Informatics*, 85:56–67, 2018. doi: 10.1016/j.jbi.2018.07.018.

Food and Agriculture Organization. *Manual on livestock disease surveillance and information systems*. FAO, 1999.

O. S. Idowu and O. O. Babalobi. Animal health management perspectives of rural livestock farmers in Southwest Nigeria: The place of community-based animal health workers. Technical report, Department of Veterinary Public Health and Preventive Medicine, University of Ibadan, Nigeria, 2020.

I. Jauregi Unanue and M. Piccardi. Pretrained language models and backtranslation for English-Basque biomedical neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 826–832, Online, 2020. Association for Computational Linguistics.

Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In Colin Cherry, Angela Fan, George Foster, Gholamreza (Reza) Haffari, Shahram Khadivi, Nanyun (Violet) Peng, Xiang Ren, Ehsan Shareghi, and Swabha Swayamdipta (eds.), *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pp. 126–135, Hybrid, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deeplo-1.14. URL https://aclanthology.org/2022.deeplo-1.14/.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2023. URL https://arxiv.org/abs/2211.01786.

NLLB Team, M. Ruiz Costa-jussà, J. Cross, et al. No language left behind: Scaling human-centered machine translation. arXiv, 2022.

Lucas Nunes Vieira. *Post-Editing of Machine Translation*, pp. 319–335. 08 2019.

E. Nwafor and A. U. Andy. A survey of machine translation tasks on Nigerian languages. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2022.

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.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, sep 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

S. Ranathunga, E. S. A. Lee, M. P. Skenduli, R. Shekhar, M. Alam, and R. Kaur. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):229:1–229:37, 2022. doi: 10.1145/3567592.

A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov. Low-resource neural machine translation improvement using source-side monolingual data. *Applied Sciences*, 13(1201), 2023. doi: 10.3390/app13021201.

Sarah Valentin, Elena Arsevska, Sylvain Falala, Jocelyn de Goër, Renaud Lancelot, Alizé Mercier, Julien Rabatel, and Mathieu Roche. Padi-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169:105163, 2020. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2019.105163. URL https://www.sciencedirect.com/science/article/pii/S0168169919310646.

Luv Verma and Ketaki N. Kolhatkar. Optimizing transformer-based machine translation model for single gpu training: a hyperparameter ablation study, 2023. URL https://arxiv.org/abs/2308.06017.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages, 2024. URL https://arxiv.org/abs/2311.09828.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021. URL https://arxiv.org/abs/2010.11934.

Xinyi Yang, Runzhe Zhan, Derek F. Wong, Junchao Wu, and Lidia S. Chao. Human-in-the-loop machine translation with large language model. In Masaru Yamada and Felix do Carmo (eds.), *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pp. 88–98, Macau SAR, China, sep 2023. Asia-Pacific Association for Machine Translation. URL https://aclanthology.org/2023.mtsummit-users.8/.

M. Zappatore and G. Ruggieri. Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language*, pp. 101582, 2023.

B. Zoph, D. Yuret, J. May, and K. Knight. Transfer learning for low-resource neural machine translation. 2016.