ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL

Volume 8 (Issue 1) (2024) Pages 64-69

Machine Translation for Improved Access to Healthcare Information in Remote Zambian Communities

Victor Neene
Department of Computer Science
ZCAS University
Lusaka, Zambia
victor.neene@zcasu.edu.zm

Douglas Kunda
Department of Computer Science
ZCAS University
Lusaka, Zambia
douglas.kunda@zcasu.edu.zm

***Abstract*—Language barriers pose a significant obstacle to achieving national health goals in Zambia. Patients struggle to understand critical medical information that in the process hinder diagnosis, treatment and preventive care. This research proposes a groundbreaking approach to bridge this gap by developing a comprehensive parallel medical linguistic corpora resource and a specialized Machine Translation (MT) system tailored to Zambia's diverse languages. Zambia's Eighth National Development Plan prioritizes improved healthcare outcomes. However, widespread language barriers between patients, healthcare providers and public health initiatives create communication gaps. Costly interpreter services are often unavailable, leaving vulnerable populations without access to crucial medical information. This is particularly concerning for mothers seeking maternal care, communities battling disease outbreaks and researchers struggling to include diverse populations in medical studies. This study will address these challenges by developing a next generation MT system and a parallel medical corpus specifically designed for Zambian Low Resourced Languages(LRL). Additionally, the study will pioneer new evaluation metrics tailored to the specific needs of medical translations. The MT system, accessible through a user friendly mobile application, will empower Zambians to access vital health information in their native languages. This will improve communication between patients and healthcare providers that will lead to better diagnoses, treatment plans and overall health outcomes. Furthermore, the research will contribute valuable insights to the broader field of MT, advancing the technology for low-resource languages and specialized domains worldwide.**

Keywords—Data Augmentation, Few Shot Learning, Resources Languages, Natural Language Processing, Parallel Corpus, Zero Short Learning

I. I. INTRODUCTION

The Eighth National Development Plan (2022–2026) highlights improving health, food and nutrition outcomes as a key development goal [1]. However, language barriers between patients and healthcare providers hinder the quality and safety of care. Larger institutions address this issue through interpreters and improved communication methods although these solutions can be both costly and time-consuming [2]. Regional language choices are made to maximize the reach of health information [3] yet in developing countries like Zambia, many patients cannot access health information in their native languages [4]. There is a significant mismatch between the languages spoken by the majority of the population and those used in healthcare services[5] resulting in communication challenges, misdiagnoses and increased patient anxiety [6].

Language barriers also negatively impact mothers seeking health information and have been associated with instances of verbal abuse during childbirth [7][8]. Community Health Workers (CHWs) in Zambia face difficulties due to language differences which affect their ability to deliver services effectively [9]. Researchers also encounter challenges with translation services when working with diverse languages which can lead to the exclusion of certain populations from medical research [10]. Machine Translation (MT) offers a potential solution to bridge these language gaps. The MT market, valued at USD 978.2 million in 2022, is growing rapidly [11]. However, MT faces challenges when dealing with Low-Resourced Languages (LRLs) such as Zambian languages particularly due to limited data availability and domain-specific translation issues [12] [13]. Moreover, robust evaluation methods for MT systems are lacking, which impedes progress [14]. MT systems also struggle with domain-specific terminology and can suffer from overfitting and brittleness [15] [16]. Despite MT's potential, its adoption in healthcare remains limited due to early-stage testing, concerns about reliability and the absence of standardized evaluation methods [17] [18].

The aim of this research is to enhance access to healthcare information in Zambia by developing a comprehensive linguistic resource and a specialized translation system. This system will be rigorously evaluated for effectiveness and deployed within a mobile application to facilitate the widespread dissemination of localized medical knowledge.

The specific research objectives are as follows:

1. To compile a parallel corpus of English medical text and its corresponding translations in target Zambian languages.

2. To develop an MT system tailored for translating healthcare information into local Zambian languages.
3. To evaluate the system's performance using metrics such as BLEU, TER and human evaluation.
4. To develop a mobile application that can be integrated with the MT model.

II. Literature Review

A. *Neural MT (NMT) and Its Challenges*

Neural MT (NMT) has significantly improved translation quality by learning effective sentence representations. A notable innovation is the multilingual model featuring an intermediate cross-lingual layer, the attention bridge, which generates language-agnostic sentence representations for transfer learning. The performance of these models varies based on the size of the bridge: larger layers enhance translation quality and trainable tasks while smaller ones improve non-trainable tasks through increased compression. Although these models benefit trainable downstream tasks, additional language signals do not enhance non-trainable benchmarks [19].

Attention-based NMT models have advanced translation efficiency, particularly for low-resource languages, by focusing on the relevant parts of the source sequence. These models manage limited parallel corpora effectively, offering improved translation accuracy and performance [20].

However, the performance of NMT on low-resource languages remains suboptimal due to limited parallel data prompting researchers to develop techniques to address this challenge [21].

B. *MT Techniques and Enhancements*

The Transformer network has become pivotal in MT, effectively capturing long-range dependencies. Its success is attributed to advances in hardware, larger datasets and sophisticated word embeddings. Transfer learning, which leverages pre-trained models, has also transformed MT. However, it often requires extensive labeled data, which can be a challenge in resource-constrained settings [22].

Transformers and Luong attention mechanisms in LSTM architectures are widely used with the BLEU score serving as a common evaluation metric. Tools like OpenNMT and Fairseq, along with corpora such as IWSLT and WMT, are standard in model development [12].

C. *Data Scarcity and Data Augmentation*

While deep learning has significantly improved MT, large datasets are essential to avoid overfitting. Data scarcity, especially for low-resource languages, remains a critical challenge. Data augmentation techniques can artificially expand datasets, enhancing model accuracy and robustness while reducing data collection costs [23]. Researchers are also exploring methods to leverage monolingual data to improve NMT for low-resource languages [24].

D. Few-Shot Learning and Zero-Shot Learning

Few-shot learning (FSL) aims to train models with minimal data, making it particularly beneficial for low-resource languages. Techniques such as meta-learning and transfer learning are being explored to improve the effectiveness of FSL. Zero-shot learning allows models to classify novel classes not encountered during training, offering potential for identifying new categories [25].

E. *Human-in-the-Loop Machine Learning*

Human-in-the-loop (HITL) machine learning integrates human expertise into the process, enhancing model performance and tackling complex tasks. HITL approaches improve accuracy by combining human insights with machine learning and they can be applied in data processing, model training and system-independent design [26].

F. *Challenges in MT*

Despite advancements, MT still faces challenges such as word alignment issues and computational demands. Problems like out-of-vocabulary terms and morphologically rich words can negatively impact translation quality. Techniques such as morphological segmentation and tokenization can help improve outcomes [27]. In multilingual settings, MT is further complicated by the diversity of grammar, syntax and semantics across languages [28]. Semantic Web technologies may assist in resolving translation ambiguities, though further development is needed [29].

G. *Application of MT in Healthcare*

Language barriers in healthcare can be mitigated by MT, particularly in emergency situations or where other options are unavailable. MT has supported refugee healthcare, though it faces challenges such as the digital divide [17]. Some studies have reported successful MT use for medical texts but rigorous evaluation is still required. Training, human editing and access to large text corpora enhance MT quality [30].

However, standardized medical terminologies tend to be English-centric, limiting global research efforts. MT methods like SMT and NMT help translate these terms [31]. While tools like Google Translate and DeepL are useful, they often require human post-editing for accuracy. Further research is needed to explore the impact of MT in interactive healthcare scenarios [32]. Improved MT tools can enhance communication between patients and providers ultimately improving healthcare outcomes [33]. Despite MT's potential, professional interpreters remain crucial for ensuring accurate communication [34].

H. *Gap Analysis*

1. The absence of parallel data for medical vocabulary hinders translation accuracy. Such data is essential for training MT models to ensure consistent and accurate translations of medical terms.
2. The lack of specialized MT models tailored to the medical context and vocabulary of Zambian LRLs poses a significant barrier. Custom models are

crucial for improving translation accuracy and usability.

3. Developing high-quality synthetic medical data is vital. Techniques to generate synthetic data will expand the training corpus, improving the robustness of MT models.
4. There is a need for new evaluation metrics focused on medical accuracy, patient comprehension and cultural appropriateness. Traditional metrics do not adequately address these aspects.
5. Integrating human expertise with MT models is essential. A human-in-the-loop approach that leverages local language proficiency can improve translation quality.
6. Evaluating MT tools in real-time healthcare communication settings is crucial. This will help identify strengths and weaknesses, guiding future improvements for clear and accurate communication in clinical environments.

III. III. Methodology

The methodology for compiling a parallel corpus of English medical text and its translations into Zambian languages will involve several steps to ensure high-quality data collection, preprocessing and quality control.

A. *Data Collection*

Partnerships will be established with hospitals, clinics and healthcare providers in Zambia to collect anonymized medical documents and patient education materials that already have translations into the target languages. Public Medical Datasets: The research will explore publicly available medical datasets from international health organizations and government initiatives, with a focus on datasets containing English medical text and corresponding translations in the target languages [35].

A platform will be developed to crowdsource translations from qualified medical professionals fluent in English and the target languages, with expertise in relevant medical domains. To ensure data accuracy and consistency, quality control measures will be put in place [36].

B. *Data Preprocessing*

Irrelevant information, such as headers, footers, formatting inconsistencies and personally identifiable information (PII) will be removed to maintain privacy.

Sentences will be tokenized into individual words or units and medical terminology will be standardized to ensure consistency across documents.

English sentences and their translations in the target languages will be aligned using software or manual processes to create accurate training data for the machine translation (MT) models.

C. Quality Control

Independent medical professionals will review and validate crowdsourced translations to ensure accuracy [36].

Terminology Consistency Checks: Terminology lists and glossaries specific to the Zambian medical context will be developed to ensure consistent translation of medical terms.

A bilingual team will review a random sample of the corpus to identify and correct any errors or inconsistencies.

D. *Corpus Annotation*

Based on research goals, additional annotation layers such as Part-of-Speech (POS) tagging [37], Named Entity Recognition (NER)[38] and domain-specific terminology tagging [39] may be incorporated to enrich the corpus.

E. *MT Model Selection and Training*

Neural MT (NMT) architectures suitable for low-resource settings will be evaluated, taking into consideration factors such as model complexity and resource requirements [40]. The selected model will be trained on the parallel corpus, with the potential use of transfer learning to enhance performance with healthcare-specific data.

F. *Domain Adaptation*

Domain adaptation techniques, such as incorporating medical dictionaries and ontologies, will be employed to improve the model's ability to handle medical terminology [15] [16].

G. *Evaluate the System's Performance Using Metrics Like BLEU, TER and Human Evaluation*

A comprehensive evaluation will be conducted using both automatic metrics (BLEU, TER) [41] and human evaluation to assess translation quality in terms of fluency, adequacy and cohesiveness [42]. This combined analysis will provide insights into the model's strengths and areas for improvement.

H. *Ethical Issues*

This project involves several ethical considerations:

An unrepresentative training dataset may result in biased translations that disadvantage certain demographics or medical conditions. To address this, the project will prioritize creating a diverse and inclusive corpus reflective of the Zambian population.

Translating medical information without considering cultural context can lead to misunderstandings and harm. To mitigate this risk, Zambian healthcare professionals and community members will be involved to ensure translations are accurate and culturally sensitive.

Sole reliance on MT for medical communication may endanger patient safety. Therefore, the project will ensure the app serves as a supplementary tool with human oversight in critical situations.

H. *Proposed Conceptual Framework*

The project envisions a machine translation system for Zambian healthcare. English medical texts will be collected alongside their Zambian language translations, cleaned and used to build a high-quality training dataset. A machine translation model will be selected, trained and fine-tuned for healthcare-specific needs. The model will be evaluated and integrated into a user-friendly mobile application for real-time translation, with potential offline functionality. The goal is to deliver healthcare information in local languages, improving health communication outcomes.

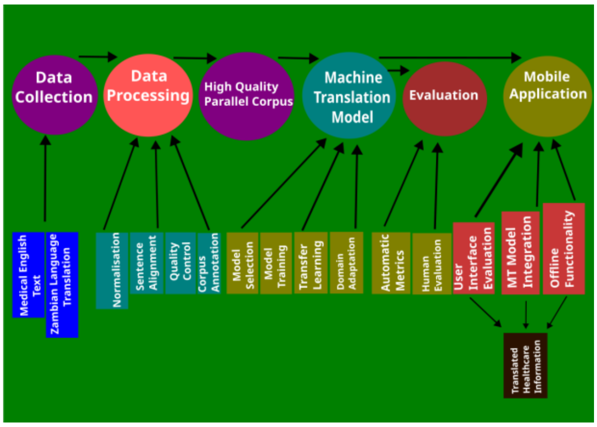

Figure 3.1: Conceptual Model

IV. Contribution and Implications

A. Contributions

This research addresses machine translation (MT) for Zambia's diverse linguistic landscape, with a focus on three main contributions:

The study will develop specialized corpora of medical texts in Zambian languages to serve as foundational training data for MT models. This will improve translation accuracy and fluency, ultimately enhancing access to healthcare services in local languages.

The research will enhance MT models for Zambia's low-resource languages (LRLs), exploring techniques to utilize existing linguistic resources and gather new data. It will tackle the unique challenges of translating these languages to improve communication in healthcare.

New evaluation metrics specific to medical translations in LRLs will be developed. These metrics will capture the nuances of medical language to ensure accurate assessments and higher translation quality.

B. Implications

This research has the potential to revolutionize healthcare access in Zambia by improving MT technology, advancing translation research for LRLs and empowering healthcare professionals and patients through enhanced communication.

V. Conclusion

This research aims to transform healthcare access in Zambia by developing specialized medical corpora and next-generation machine translation (MT) models for low-resource languages (LRLs). These models will deliver accurate and fluent medical translations, improving communication and information dissemination. The study will also introduce new evaluation metrics tailored to medical translations in LRLs, leading to higher standards of accuracy and usability. The impact will be significant: Zambians will gain access to vital health information in their native languages, enabling them to make informed decisions while healthcare professionals will benefit from improved communication with patients, leading to better diagnosis and treatment. Moreover, the research will contribute to advancing MT technology for LRLs globally.

References

1] G. of the Repblic of Zambia, "The eighth national development plan,"

[2] H. Al Shamsi, A. G. Almutairi, S. Al Mashrafi and T. Al Kalbani, "Implications of language barriers for healthcare: a systematic review," Oman medical journal, vol. 35, no. 2, p. e122, 2020.

[3] M. Civico, "Covid-19 and language barriers," Çalı¸sma Notu, no. 21-4, 2021.

[4] O. D. Obaremi and W. M. Olatokun, "A survey of health information source use in rural communities identifies complex health literacy barriers," Health Information & Libraries Journal, vol. 39, no. 1, pp. 59–67, 2022.

[5] K. Batchelor, L. A. Yoda, F. E. G. S. Ouattara and O. Hellewell, "and strategic planning for hiv/aids-related health care and communication," Wellcome Open Research, vol. 4, 2019.

[6] B. M. Sichlindi, H. Mulunda, A. Chishiba, F. Simui, et al., "Disablers to access to healthcare services experienced by learners with hearing impairment at musakanya school in mpika district, zambia," International Journal of Research and Scientific Innovation, vol. 9, no. 1, pp. 65–74, 2022.

[7] F. Mulauzi, D. Chisha, I. Alwisho, E. Chiumia, C. Makasa and D. Hakalumbwe, "Health information literacy, information needs and in formation seeking behaviours among mothers with children under the age of five: a case study of chilenje level one hospital in lusaka, zambia.," 2020.

[8] C. Mweemba, M. Mapulanga, C. Jacobs, P. Katowa-Mukwato and vM. Maimbolwa, "Access barriers to maternal healthcare services in selectedvhard-to-Pan African Med-ical Journal, vol. 40, no. 1, 2021.

[9] G. E. Khumalo, E. E. Lutge, P. Naidoo and T. P. Mashamba-Thompson,v"Barriers and facilitators of rendering hiv services by community healthvworkers in sub-

saharan africa: a meta- synthesis," Family Medicine and Community Health, vol. 9, no. 4, 2021.

[10] M. Mukosha, D. Muyunda, S. Mudenda, M. K. Lubeya, A. Kumwenda,vL. M. Mwangu and P. Kaonga,v"Knowledge, attitude and practice towards cervical cancer screening among women living with human immunodeficiency virus: Implication for prevention strategy uptake," Nursing Open,vvol. 10, no. 4, pp. 2132–2141, 2023.

[11] G. V. Research, "Machine translation market size and trends." https://www.grandviewresearch.com/industry-analysis/machine-translation-market, 2023.

[12] B. K. Yazar, D. Ö. ¸Sahın and E. Kiliç, "Low-resource neural machine translation: A systematic literature review," IEEE Access, vol. 11, pp. 131775131813, 2023.

[13] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl and A. Birch, "Surveyvof low-resource machine translation," Computational Linguistics, vol. 48,vno. 3, pp. 673–732, 2022.

[14] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan,vM. Ranzato, F. Guzmán and A. Fan, "The flores-101 evaluation benchmarkvfor low-resource and multilingual machine translation," Transactions of thev Association for Computational Linguistics, vol. 10, pp. 522–538, 2022.

[15] C. Chu and R. Wang, "A survey of domain adaptation for machine translation," Journal of information processing, vol. 28, pp. 413–426, 2020.

[16] D. Saunders, Domain adaptation for neural machine translation. PhD thesis,2021.

[17] L. N. Vieira, M. O'Hagan and C. O'Sullivan, "Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases," Information, Communication & Society, vol. 24, no. 11, pp. 1515–1532, 2021.

[18] M. Pluymaekers, "How well do real-time machine translation apps perform in practice? insights from a literature review," in Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pp. 51–60, 2022.

[19] R. Vázquez, A. Raganato, M. Creutz and J. Tiedemann, "A systematic study of inner-attention-based sentence representations in multilingual neural machine translation," Computational Linguistics, vol. 46, no. 2, pp. 387–424, 2020.

[20] M. Koivuniemi, "Attention-based neural machine translation: a systematic mapping study," 2020.

[21] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam and R. Kaur, "Neural machine translation for low-resource languages: A survey," ACM Computing Surveys, vol. 55, no. 11, pp. 1–37, 2023.19

[22] E. Kotei and R. Thirunavukarasu, "A systematic review of transformer-based pre-trained language models through self-supervised learning," Information, vol. 14, no. 3, p. 187, 2023.

[23] M. A. Bansal, D. R. Sharma and D. M. Kathuria, "A systematic review on data scarcity problem in deep learning: solution and applications," ACM Computing Surveys (CSUR), vol. 54, no. 10s, pp. 1–29, 2022.

[24] S. M. U. Qumar, M. Azim and S. Quadri, "Neural machine translation: A survey of methods used for low resource languages," in 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1640–1647, IEEE, 2023.

[25] W. Wang, V. W. Zheng, H. Yu and C. Miao, "A survey of zero-shot learning: Settings, methods and applications," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 10, no. 2, pp. 1–37, 2019.

[26] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma and L. He, "A survey of humanin-the-loop for machine learning," Future Generation Computer Systems, vol. 135, pp. 364–381, 2022.

[27] S. A. ALMAAYTAH and S. A. ALZOBIDY, "Challenges in rendering arabic text to english using machine translation: A systematic,"

[28] Sitender, S. Bawa, M. Kumar and Sangeeta, "A comprehensive survey on machine translation for english, hindi and sanskrit languages," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 4, pp. 3441– 3474, 2023.

[29] D. Moussallem, M. Wauer and A.-C. N. Ngomo, "Machine translation using semantic web technologies: A survey," Journal of Web Semantics, vol. 51, pp. 1–19, 2018.

[30] M. Zappatore and G. Ruggieri, "Adopting machine translation in the healthcare sector: A methodological multi-criteria review," Computer Speech & Language, p. 101582, 2023.

[31] R. Noll, L. S. Frischen, M. Boeker, H. Storf and J. Schaaf, "Machine translation of standardised medical terminology using natural language processing:A scoping review," New Biotechnology, 2023.

[32] P. S. Herrera-Espejel and S. Rach, "The use of machine translation for out reach and health communication in epidemiology and public health: Scoping review," JMIR Public Health and Surveillance, vol. 9, no. 1, p. e50814,2023.

[33] A. Kreienbrinck, S. Hanft-Robert and M. Mösko, "Usability of technological tools to overcome language barriers in health care: a scoping review protocol," BMJ open, vol. 14, no. 3, p. e079814, 2024.

[34] P. Hudelson and F. Chappuis, "Using voice-to-voice machine translation to overcome language barriers in clinical communication: An exploratory study," Journal of General Internal Medicine, pp. 1–8, 2024.

[35] L. Oakden-Rayner, "Exploring large-scale public medical image datasets," Academic radiology, vol. 27, no. 1, pp. 106–112, 2020.

[36] C. Ye, J. Coco, A. Epishova, C. Hajaj, H. Bogardus, L. Novak, J. Denny, Y. Vorobeychik, T. Lasko, B. Malin, et al., "A crowdsourcing framework for Science Proceedings, vol. 2018, p. 273, 2018.

[37] A. Chiche and B. Yitagesu, "Part of speech tagging: a systematic review of deep learning and machine learning approaches," Journal of Big Data, vol. 9, no. 1, p. 10, 2022.

[38] J. Li, A. Sun, J. Han and C. Li, "A survey on deep learning for named entity recognition," IEEE transactions on knowledge and data engineering, vol. 34, no. 1, pp. 50–70, 2020.

[39] M. Alruqimi, N. Aknin, T. Al-Hadhrami and A. James-Taylor, "Towards semantic interoperability for iot: Combining social tagging data and wikipedia to generate a domain-specific ontology," in Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Re- liable Information and Communication Technology (IRICT 2018), pp. 355–
363, Springer, 2019.

[40] F. Stahlberg, "Neural machine translation: A review," Journal of Artificial Intelligence Research, vol. 69, pp. 343–418, 2020.

[41] M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie and A. F. Martins, "Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust," in Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 46–68,2022.

[42] V. Mendonça, R. Rei, L. Coheur and A. Sardinha, "Onception: Active learning with expert advice for real world machine translation," Computational Linguistics, vol. 49, no. 2, pp. 325–372, 2023.