A Literature Review on English–isiXhosa Machine Translation in the Medical Domain

Nick Matzopoulos
Department of Computer Science
Cape Town, South Africa
mtznic006@myuct.ac.za

ABSTRACT

Machine translation (MT) has advanced significantly in recent years, primarily through the development of Neural Machine Translation (NMT) and transformer architectures. However, these improvements are unevenly distributed across languages. Low-resource languages such as isiXhosa continue to face substantial challenges due to limited parallel corpora and domain-specific terminology. This literature review explores the evolution of MT approaches, from early neural architectures to state-of-the-art transformer models, and examines their comparative effectiveness against decoder-only large language models (LLMs). The review highlights the unique challenges of isiXhosa translation, including its morphological complexity and data scarcity. It investigates promising strategies for overcoming these constraints, including multilingual training, transfer learning from related Nguni languages, and data augmentation techniques like back-translation. Additionally, the review examines domain adaptation approaches for medical translation, comparing data-centric methods such as fine-tuning and terminology-based techniques with model-centric approaches like adapter layers. A significant focus is placed on medical domain translation, where accurate systems can bridge communication gaps between healthcare providers and isiXhosa-speaking patients. The recent introduction of an English-isiXhosa medical dataset addresses both linguistic and domain challenges, opening new opportunities for developing more effective and accessible MT solutions for healthcare applications.

1 INTRODUCTION

MT systems have evolved through several distinct paradigms over the past decades. From early rule-based methods that rely on handcrafted linguistic rules [38], to data-driven statistical approaches [1], and, most recently, neural architectures using attention [35]. Each paradigm improved translation quality but also brought new challenges, especially for languages with limited parallel data.

IsiXhosa, a Bantu language spoken by millions in South Africa, is a low-resource language where the lack of large parallel corpora has held back the development of good MT systems. Its complex morphology—which consists of agglutinative structures, many noun classes, and dialectal variations—makes accurate translation difficult [30]. Adding domain-specific needs, like translating medical text, makes robust MT even harder due to specialised terminology [11]. Common domain adaptation methods like fine-tuning and back-translation help but still need in-domain data [10]. Until recently, real isiXhosa medical text datasets were very limited [36].

A breakthrough comes from Blocker et al. [34], who created an English–isiXhosa medical dataset and benchmark for testing MT in the medical domain. This dataset will help researchers train and improve models for isiXhosa medical MT [34]. It also offers a way to benchmark medical translation quality in a low-resource language setting [34].

This literature review explores how MT architectures have evolved. The discussion covers encoder–decoder architectures [28], attention mechanisms [2], and the Transformer model [35]. It then compares them to decoder-only structures like LLMs [23]. The literature review then examines the specific challenges of isiXhosa translation: limited data, complex word structure, dialect differences, and gaps in healthcare terminology [36]. Then, it moves on to highlight promising approaches like multilingual training and transfer learning that have improved African-language MT [30]. Next, the paper details domain adaptation techniques—including fine-tuning, back-translation, and lexicon induction—that help handle specialised content [37]. Finally, the review discusses Blocker et al.'s new dataset and its potential impact on MT for isiXhosa in the medical field [34].

2 NEURAL MACHINE TRANSLATION

Early MT systems were Rule Base Machine Translation (RBMT) systems, which relied on manually crafted linguistic rules and dictionaries but often produced rigid outputs [38]. In the 1990s and 2000s, Statistical Machine Translation (SMT) brought a data-driven approach, using probabilistic models that produced more fluent translations if large parallel corpora were available [1]. Around the mid-2010s, NMT introduced end-to-end neural networks [39]. NMT is capable of capturing long-range dependencies and produces more coherent results, albeit requiring extensive parallel data [39]. In low-resource scenarios like English–isiXhosa, NMT's performance can degrade sharply unless careful adaptation methods are used [1].

2.1 Encoder–Decoder Architecture

The standard Neural Machine Translation architecture is the encoder-decoder architecture proposed by Sutskever, Cho, et al. in 2014. The encoder network reads the source sentence and encodes it into a fixed-length continuous vector (called the context vector), and then a decoder network generates the target sentence from that vector [28]. The encoder, which is often a Recurrent Neural Network (RNN) like a Long Short-Term Memory (LSTM) [41] processes the source token sequence $x_1, x_2, \ldots, x_n$ into a single vector h_n, and the decoder RNN then produces an output sequence $y_1, y_2, \ldots, y_m$ conditioned on that vector [28]. This architecture is trained end-to-end to maximise the probability of the correct translation given the source [28].

Research showed that these models can learn internal representations of translation without explicit bilingual dictionaries [24].

However, a critical limitation became evident: forcing all source information into one fixed-length vector creates a bottleneck for long sentences [28]. The encoder's last hidden state h_n must represent the entire input [28]. This becomes difficult as sentence length grows—the decoder may "forget" earlier parts of a long source sentence. This limitation is exaggerated for long inputs, where the fixed-length context vector cannot retain all semantic and syntactic details, which leads to a decline in translation accuracy [24]. Cho et al. (2014) explored this issue, noting that performance decreases significantly beyond a certain sentence length due to the compression constraints of the architecture [24].

2.2 Encoder–Decoder with Attention

To address this, Bahdanau et al. (2015) introduced the attention mechanism, which allows the decoder to attend to different parts of the source sequence at each generation step [2]. Rather than using one fixed-length context vector, like in the traditional encoder-decoder architecture, the model creates a sequence of source annotations and dynamically aligns to them during the decoding process [2]. At each target word output, the decoder computes attention weights over all source encoder states [2]. This enables the model to learn which source words are most relevant to the current translation context.

To clarify how this works, the attention mechanism computes a context vector c_i for each target word y_i as a weighted sum of the encoder's hidden states [2]:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

Here, h_j represents the encoder's hidden state for the j-th source token, and α_{ij} are the attention weights, which determine the importance of each source position j for generating the target word at position i [2]. These weights are calculated using a softmax function [2]:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

where e_{ij} is a score function (often a small neural network) that measures the relevance of source position j to target position i. This dynamic weighting allows the model to focus on different parts of the input sentence as needed, rather than relying on a single fixed vector [2]. This fixed the bottleneck problem and helped the model better handle word alignment and structure differences, which is vital for language pairs with divergent structures like English and isiXhosa [25]. Luong et al. (2015) expanded on this, showing that attention enhances translation quality by better capturing dependencies across varied word orders, a key factor in morphologically complex languages [25].

This sequence-to-sequence with attention model was a breakthrough: it solved the fixed-vector bottleneck by giving the decoder a context vector that varies for each target word, computed as a weighted sum of encoder outputs [2]. After incorporating attention, NMT models not only matched state-of-the-art phrase-based SMT on some tasks but often surpassed them, all within a single integrated model [2].

Another challenge for neural models is handling rare or unseen words. This was solved using subword modelling [3]. Sennrich et al. (2016) suggested using subword units like Byte-Pair Encoding (BPE) to break uncommon words into smaller pieces, which the model can combine to form unknown words [3]. This creates an open vocabulary: the NMT system's vocabulary includes word parts (such as common stems, prefixes, suffixes), so even a new word can be made by joining known subwords [3]. For example, the English word "internationalization" might be split into *international* + *ization*, and an isiXhosa word like *ukwabelana* (sharing) might be split into *ukwabele* + *na*. By working with subwords, NMT can learn to translate parts of words and handle word creation and combining. This was a key improvement—it greatly enhanced translation of languages with complex word structures like isiXhosa [3].

2.3 Transformer Architecture and Advanced NMT Models

A major leap in NMT came with the introduction of the Transformer model by Vaswani et al. (2017), which moved away from recurrent networks entirely [29]. The Transformer uses self-attention mechanisms to process sequences, achieving the same goal as RNNs but with better efficiency [26]. As illustrated in Figure 1 [29], the Transformer encoder consists of multiple layers that each contain a self-attention sub-layer followed by a feed-forward neural network sub-layer [29]. The decoder similarly has multiple layers, each with self-attention, an encoder–decoder attention sub-layer, and a feed-forward sub-layer [29].

This architecture allows the model to attend to any position in the input or output sequence from any other position, thanks to the self-attention mechanism [26]. Unlike an RNN, the Transformer does not process tokens sequentially—it processes the entire sequence in parallel, computing attention scores between all pairs of positions in a layer [29]. This removes the step-by-step processing of RNNs and allows for faster training and better scaling [29]. By processing in parallel, Transformers excel at handling connections between distant words, leading to better translations [26]. Popel and Bojar (2018) backed this up, showing that Transformers regularly beat RNN-based models on tasks needing broad context [26].

The impact of the Transformer on MT was immediate and significant. Transformers achieved state-of-the-art results on benchmarks, often improving BLEU scores over the best RNN models while training faster due to parallelisation [29]. For instance, on English–German translation, the Transformer significantly outperformed older LSTM-based models [29]. The ability to train on very large datasets with high parallel throughput meant researchers and companies could scale NMT models to unprecedented sizes [26]. This led to massive multilingual NMT models such as Facebook's M2M-100 [4] and NLLB-200 [4], which employ the Transformer architecture to handle hundreds of languages. NLLB-200 (2022) uses a 54-billion-parameter Transformer trained on 200 languages and achieved an average +44% BLEU improvement on low-resource language pairs compared to prior methods [4]. Such multilingual models are especially beneficial for low-resource languages like isiXhosa, where transfer learning from high-resource languages can compensate for scarce parallel data, enhancing translation quality [27]. Fan et al. (2021) highlighted similar successes with multilingual models for African languages, underscoring their potential for languages like isiXhosa [27].

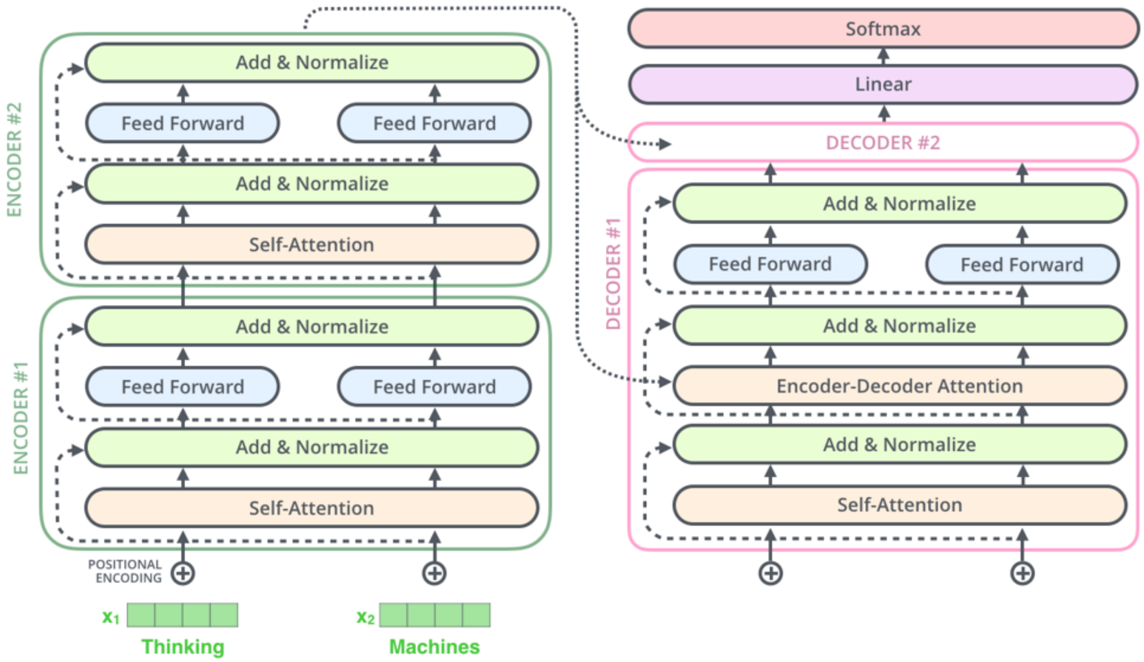

Figure 1: A schematic illustration of the Transformer architecture, showing self-attention and feed-forward layers in both encoder (green) and decoder (red). The decoder includes an encoder–decoder attention sub-layer (orange) to attend to encoder outputs. Positional encoding (PE) is added to inform the model of token positions [29].

Furthermore, these advanced architectures can be fine-tuned on domain-specific corpora, such as medical texts, to ensure precise and contextually appropriate translations in fields like healthcare [11].

2.4 2.3 Decoder–Only Models for Translation

A recent topic of interest is the role that LLMs play in translation, and how decoder-only architectures, such as ChatGPT and LLaMA, compare to traditional encoder–decoder MT models. Encoder–decoder Transformer models are explicitly designed for sequence-to-sequence tasks: the encoder processes the source text and a separate decoder generates the target text [7]. By contrast, decoder-only LLMs are trained to predict the next token in a sequence [7]. Translation using a decoder-only LM often involves prompting the model with the source sentence (for instance, "Translate to isiXhosa:") and having it generate a continuation which is the translation.

Large decoder-only LLMs have demonstrated unexpectedly strong translation performance for high-resource languages, even without explicit parallel data [8]. Wang et al. (2023) state that GPT-4 can rival or even exceed certain dedicated MT systems in languages like French, Chinese, or Spanish. This success is attributed to the extensive multilingual text that these models see during pretraining, enabling them to produce fluent output for languages highly represented in their training sets [7].

However, decoder-only LLMs show notable shortcomings in low-resource scenarios [23]. Robinson et al. (2023) present a wide-ranging evaluation of ChatGPT (GPT-3.5) against Meta's NLLB [4] on FLORES-101 [40], encompassing numerous African languages [23]. According to Robinson et al., ChatGPT's translation quality correlates strongly with the resource level of the language. Specifically, ChatGPT performed comparably to specialised MT for high-resource languages but scored significantly lower for African languages with minimal training data, such as isiXhosa [23]. In zero-shot translation, ChatGPT outperformed the NLLB system only in around 6% of the least-resourced language directions, but did so in around 47% for high-resource directions [23]. The average BLEU scores on African languages remained substantially behind NLLB, reflecting difficulties with morphology and vocabulary in low-data conditions [23].

On the other hand, encoder–decoder MT models can be specifically trained or fine-tuned [12] on parallel data for a low-resource language, thereby learning domain-specific terms and morphological patterns more effectively [23]. Thus, for isiXhosa, a focused MT approach often has a clear advantage over LLM translation. Overall, LLMs like ChatGPT and LLaMA are not yet a substitute for specialised MT in challenging low-resource scenarios.

3 ISIXHOSA IN NEURAL MACHINE TRANSLATION

IsiXhosa is a low-resource language facing significant challenges for MT systems. The primary obstacle is the scarcity of labelled data for English–isiXhosa—where tens of millions of parallel sentences are common for languages like English–French, but only tens of thousands exist for English–isiXhosa [5]. This limited data leads to poor translations that either omit content or retain untranslated English text [5]. Recent evaluations reveal that while MT performance is improving, low-resource languages like isiXhosa continue to lag significantly behind better-resourced African languages [6]. This data scarcity is compounded by isiXhosa's morphological complexity: it is highly agglutinative, with a range of grammatical features in a single word [6]. As a result, models must handle a vast number of word forms that may never appear explicitly in a small training set [5].

3.1 Multilingual NMT

One effective approach for mitigating data scarcity in isiXhosa MT is **multilingual NMT**, where a single model is trained on multiple languages to enable cross-lingual transfer [30]. Recent advances in massively multilingual models like NLLB have enabled significant gains for low-resource translations [32]. For instance, Emezue et al. built "MMTAfrica," the first many-to-many NMT system for several African languages including isiXhosa [31]. By training a Transformer on a combined corpus of six African languages plus English and French, Emezue et al. leveraged shared language representations to improve translation quality for each low-resource language. Notably, the multilingual model outperformed a strong bilingual baseline on all test directions, with the largest gain observed on isiXhosa: a French→isiXhosa translation improved by over +19 spBLEU compared to the prior baseline [31].

Nyoni and Bassett likewise reported that a trilingual English–isiXhosa–isiZulu model achieved the best performance among various setups, surpassing both transfer learning and zero-shot methods on English→isiZulu translation [30]. In fact, the trilingual model boosted BLEU by +9.9 points over a baseline system, more than doubling the previous state of the art [30]. These findings demonstrate that multilingual training can significantly improve isiXhosa MT by sharing strength across related languages.

Multilingual NMT was also validated in the WMT22 Africa Translation Task, where top systems used single models for English and multiple African languages. The UCT WMT22 submission trained one Transformer on English and eight South/South-East African languages, as well as certain language–pair directions (e.g. isiXhosa↔Zulu) [33]. The model can translate between English and all included languages, and even between African languages themselves, showing the feasibility and value of broad multilingual MT for isiXhosa. Overall, multilingual NMT provides a data-driven way to transfer generalisation from higher-resource or related sister languages to isiXhosa, improving translation quality in low-resource languages [30].

3.2 Transfer Learning

Transfer learning helps solve the problem of limited parallel data for isiXhosa by starting the translation model with knowledge from other translation tasks [30]. A popular method is to first train an NMT model on a high-resource or related language pair, then tune it for English–isiXhosa [30]. Research shows that transfer learning works best when using a language closely related to isiXhosa [30]. Nyoni and Bassett showed this by using English–isiXhosa as a starting point for English–isiZulu: beginning with an English–isiXhosa model and tuning it on a smaller English–isiZulu dataset improved results by +6.1 BLEU compared to starting from scratch [30]. In contrast, using an unrelated language (English–Shona) gave only small improvements [30]. Using an African language model (particularly one from the same language family) as the transfer source often gives better results than using an English-based model [31].

Another type of transfer learning is fine tuning large pre-trained multilingual models. For example, Emezue et al. tuned a 580M-parameter multilingual NMT model on their African translation dataset [31] which can be adapted to specific domains. Overall, transfer learning—whether from related-language models or multilingual pre-training—has been essential for isiXhosa MT, as long as the source model shares key language features with isiXhosa [30, 31].

3.3 Zero-Shot Learning

When parallel data for a language pair is completely unavailable, zero-shot learning translates purely via a model's cross-lingual capability [30]. Zero-shot MT for isiXhosa might rely on training a many-to-many model on other language pairs (e.g. English–Zulu, English–Swahili), then asking it to handle English→isiXhosa. Early experiments showed limited yet promising results for zero-shot isiXhosa translation. For instance, an English–isiXhosa–isiZulu model generated good English→isiZulu output even with zero parallel Zulu examples, improving BLEU by a modest +2.0 [30]. However, performance was still significantly lower than models trained with direct data or strong transfer sources [30]. In general, pure zero-shot English–isiXhosa translation remains challenging, given the pair's data scarcity.

Modern multilingual NMT systems do facilitate zero-shot capabilities: the UCT WMT22 system was trained on many directions and could implicitly translate between isiXhosa and Zulu despite no direct corpus [33]. Thus, zero-shot is a fallback option where no direct data is available, though it usually benefits from subsequent data augmentation or fine-tuning [30].

3.4 Data Augmentation Strategies

Given the limited authentic data for isiXhosa, various data augmentation strategies can create additional training examples or adapt models more effectively. Two widely used approaches are *fine-tuning* and *back-translation.*

3.4.1 Fine-Tuning.

Fine-tuning a pretrained model on a small English–isiXhosa parallel set is standard in low-resource MT [30]. One can start with a multilingual model or a related-language model, then refine it on isiXhosa. This gives substantial gains in translation accuracy, as the model focuses on isiXhosa's specific morphology once it has a strong general initialisation [31]. In specialised domains, like the medical domain, researchers might pretrain on general data and then fine-tune on a small in-domain isiXhosa corpus [30]. Past work shows that fine-tuning large models for isiXhosa often outperforms training them from scratch [30].

3.4.2 Back-Translation.

Back-translation (BT) is where monolingual target-language data is translated back into the source language to create synthetic parallel sentences. [30] For English–isiXhosa, one would translate isiXhosa monolingual text into English, then add these pairs (synthetic English, original isiXhosa) to the training data for English→isiXhosa. This approach has proved highly effective in low-resource MT [33], with Emezue et al. citing back-translation as a key factor in boosting BLEU on the FLORES [40] test set [32]. The UCT WMT22 system used BT to generate synthetic translations for under-covered pairs (like isiXhosa↔Zulu) and saw marked improvements [33].

3.5 Datasets and Evaluation Benchmarks

A major resource for evaluating isiXhosa MT is the FLORES-101 [40] test benchmark, covering 101 languages (including isiXhosa). Tasks such as WMT22 have used FLORES dev/test sets for consistent comparisons. The best WMT22 systems reached around 38 BLEU on isiXhosa↔English, a substantial jump from older baselines [32]. Although still far behind high-resource languages, these methods show rapid progress in African MT.

IsiXhosa, being the low-resource language that it is, leads to researchers heavily relying on the techniques described above—especially multilingual training, transfer learning, and back-translation—to achieve viable results for isiXhosa MT.

4 DOMAIN ADAPTATION FOR NEURAL MACHINE TRANSLATION

To improve MT performance for specific domains or for low-resource languages like isiXhosa, researchers employ various domain adaptation techniques [10]. These methods aim to adapt a general MT model, trained on out-of-domain or general data, to a specific domain [10]. This section will discuss the two main categories of domain adaptation: data-centric and model-centric methods [10]. Data-centric methods primarily generate domain-specific data to fine-tune the model, while model-centric methods modify the Machine Translation architecture [12]. First, the challenges in isiXhosa domain adaptation will be discussed.

4.1 Challenges in isiXhosa Domain Adaptation

Translating domain-specific content—particularly medical text—into isiXhosa introduces further challenges. The medical domain is terminology-heavy and often involves specialised vocabulary that may not have direct equivalents in isiXhosa [9]. One study on translating a medical questionnaire from English to isiXhosa found significant difficulty in providing medical terms accurately [9]. For example, terms like “mobility” or phrases like “confined to bed” had no straightforward isiXhosa translations and required descriptive phrasing or were prone to misinterpretation [9]. In many cases, healthcare practitioners in South Africa resort to using English medical terms when speaking isiXhosa, because an isiXhosa term is either missing or not widely known [9]. This means an MT system might encounter an English medical word and fail to find examples in the isiXhosa training data, leading it to leave the term untranslated or guess incorrectly [30]. Domain-specific translation into isiXhosa magnifies the low-resource problem: not only is general parallel data scarce, but parallel medical data is even scarcer. An MT system needs domain adaptation to handle this content well.

4.2 Data-Centric Approaches to Domain Adaptation

Fine-tuning and synthetic data generation lie at the core of data-centric adaptation strategies. The goal is to bring the training distribution closer to the target domain by reweighting the model’s exposure to domain-relevant examples [12].

4.2.1 Fine-Tuning for Domain Adaptation. Fine-tuning is a widely used technique in NMT that adapts a pre-trained model to a specific domain [12]. It involves taking a model initially trained on a large, general-purpose dataset and further training it on a smaller, domain-specific dataset, such as medical texts for English–isiXhosa translation [13].

Synthetic data generation techniques, such as back-translation or dictionary-based augmentation, are often used to create such fine-tuning datasets in low-resource scenarios [13]. Rather than relying solely on rare in-domain parallel corpora, researchers can fine-tune models on generated text that reflect the vocabulary of the target domain [13].

4.2.2 Synthetic Data Generation. In low-resource contexts, an effective way to adapt models to a specific domain is by creating *synthetic in-domain data.* This leverages existing monolingual corpora and sometimes dictionaries or lexicons to produce parallel examples for training.

Back-Translation (BT). Back-translation is a widely used technique for low-resource and domain-specific MT [**?**]. It creates synthetic training data by using monolingual target-language text [12]. The process follows a three-step pipeline [**?**]: first, we train a reverse model (e.g. isiXhosa→English) using available parallel data; second, we use this model to translate monolingual isiXhosa sentences into English, generating synthetic English text; and finally, we pair these with the original isiXhosa lines to form new synthetic parallel data (English’, isiXhosa). This synthetic data is then used to train or fine-tune the forward model (English→isiXhosa).

Importantly, the slight noise introduced during this process can help regularise the model, improving generalisation to unseen structures [13]. For domains like medicine, where authentic English–isiXhosa bitext is scarce, BT allows us to inject domain-relevant vocabulary into training [14, 15]. The UCT WMT22 team demonstrated that even modest use of BT significantly improved BLEU scores for African languages, particularly when integrated with multilingual pretraining and iterative fine-tuning [**?**].

Forward Translation (FT). Forward translation inverts the idea of back-translation: we take monolingual *source*-language sentences from the target domain and translate them into the *target* language using an existing model [12]. FT is especially useful when the source side (e.g. English) has domain-specific data that we want to incorporate into training [16]. FT is less common than BT — particularly for extremely low-resource target languages [17]. FT can still help if the baseline model is decent and domain-relevant source texts are available [17].

4.2.3 Lexicon-Based and Terminology-Focused Approaches.

DALI: Domain Adaptation by Lexicon Induction. DALI focuses on generating domain-specific bilingual lexicons to handle unknown or specialised terms [18]. It induces a lexicon from in-domain monolingual data and uses it to produce pseudo-parallel sentences [18]. For example, if we have a isiXhosa medical corpus, DALI aligns isiXhosa words with English equivalents, then replaces isiXhosa words with the induced English terms to create synthetic parallel text. Research showed large BLEU gains when no real in-domain parallel data was available [18], highlighting DALI's relevance for low-resourced scenarios like isiXhosa medical texts.

Bilingual Word Embeddings for OOV Terminology. Another tactic for domain terminology is using bilingual word embeddings to detect out-of-vocabulary (OOV) words and propose translations [19]. By aligning monolingual embeddings for English and isiXhosa, one can map specialised English terms to their nearest isiXhosa neighbours. This helps with OOV errors in domain text, by automatically substituting the correct term during training [19].

Dictionary-Based Data Augmentation (DDA). Dictionary-based augmentation inserts domain-specific terms into general parallel sentences, creating pseudo in-domain training examples [20]. For instance, if a dictionary shows "diabetes" = "iswekile" DDA might find good spots in general sentence pairs and swap common words with these special terms, creating new examples with domain words. Tests show that using DDA with fine-tuning or back-translation leads to big improvements, mainly by helping the model learn rare domain words [20].

4.2.4 *LLM-Based Domain Adaptation.*

Dictionary-Informed Prompt-Based MT (DIPMT). Prompt-based approaches use LLMs without retraining them. DIPMT puts a domain-specific word list in the prompt to guide the LLM's translation [22]. For English–isiXhosa medical text, a small dictionary of medical terms could be given to the LLM matching English terms to isiXhosa. This helps the LLM use correct domain terms [22]. Studies show DIPMT can work much better than basic prompting, especially for specialised domains [22].

Rihan et al. suggest using an LLM to create in-domain text that includes specific terms, then fine-tuning the main NMT model on this new parallel data [21]. They found big improvements in term accuracy, increasing from 36.67% to 72.88, on test sets [21]. For isiXhosa, an LLM with sufficient knowledge could create isiXhosa sentences with medical terms, filling data gaps in ways that normal data collection cannot easily do [21].

4.3 Model-Centric Approaches

While data-centric methods focus on creating or enhancing training data, model-centric techniques alter the NMT architecture or training procedure to handle domain differences [12].

Adapter Layers and Domain-Specific Embeddings. A popular method is to add domain-specific adapter layers to a pretrained model [10]. These are small networks inserted between existing layers and trained on in-domain data, while the base parameters remain unchanged [12]. Adapters let the model preserve its original capabilities yet learn new domain features [12].

Model-centric solutions may require more complex engineering than data-centric ones, but they offer fine-grained control over how much knowledge is shared across domains [12]. For instance, if we want an English–isiXhosa system to handle both everyday text and medical text, adding a medical adapter could let us activate that domain knowledge only when needed.

5 EVALUATION AND BENCHMARKS

Evaluating domain adaptation for isiXhosa demands more than a single BLEU score on general test sets. Especially in medical translation, accuracy on specialised terminology is critical. A model might achieve decent BLEU but fail at translating “confined to bed” or “mobility” [9]. The new dataset from Blocker et al. now provides researchers with a much-needed medical isiXhosa testbed, allowing for evaluation of translation quality specifically in the medical domain. This benchmark enables more targeted assessment of how well MT systems handle medical terminology and patient-doctor dialogues in isiXhosa [9].

6 DISCUSSION

The newly introduced English-isiXhosa medical dataset represents a significant advancement for machine translation research. This corpus directly addresses the persistent data scarcity challenge for isiXhosa by providing authentic, domain-specific parallel text from medical consultations. Prior research has consistently identified limited data as the primary obstacle for English-isiXhosa translation, with tens of thousands of parallel sentences available compared to tens of millions for high-resource language pairs [5]. The dataset is particularly valuable because it combines two significant challenges: isiXhosa as a morphologically complex, low-resource language and the specialised terminology of the medical domain.

Research on isiXhosa MT has shown that current models achieve modest results compared to high-resource languages, with the best WMT22 systems reaching approximately 38 BLEU for isiXhosa↔English translation [32]. Several approaches have demonstrated effectiveness in improving translation quality. Multilingual NMT, where a single model learns to translate between multiple language pairs simultaneously, is one of the leading methods, with studies showing that models trained on multiple related African languages significantly enhance isiXhosa translation performance [30]. For instance, MMTAfrica demonstrated that a multilingual model outperformed bilingual baselines, with French→isiXhosa translation improving by over 19 BLEU [31]. Similarly, Nyoni and Bassett found that trilingual English-isiXhosa-isiZulu models boosted BLEU scores by 9.9 points over baseline systems [30]. Transfer learning, which leverages knowledge from one language pair to improve another with fine-tuning on English-isiXhosa yields a 6.1 BLEU improvement for English-isiZulu translation compared to training from scratch [30]. Data augmentation through back-translation, a technique that translates monolingual text to create synthetic parallel data, has allowed researchers to make use of monolingual isiXhosa text [33].

Domain adaptation for MT represents another critical area when translating specialised content like medical text. Data-centric adaptation strategies primarily focus on bringing the training distribution closer to the target domain. Fine-tuning a general model on a smaller domain-specific dataset has proven effective for adapting to specialised medical vocabulary and stylistic patterns. Synthetic data generation techniques address data scarcity: back-translation creates artificial parallel data by translating monolingual isiXhosa text into English, with the WMT22 UCT team demonstrating significant BLEU score improvements for African languages through this approach [33].

Lexicon-based approaches specifically target domain terminology, which is crucial for medical translation. DALI generates domain-specific bilingual lexicons, showing substantial improvements when in-domain parallel data is unavailable [18]. DiPMT is a recent method that supplies domain-specific terminology in prompts to guide large language models during translation, significantly improving performance on specialised vocabulary [22].

The newly introduced English-isiXhosa medical dataset presents a unique opportunity at the intersection of low-resource language translation and domain-specific adaptation [34]. This combination creates a particularly challenging scenario: the data scarcity of isiXhosa text combined with specialised medical terminology compounds the difficulty of developing accurate translation systems. The dataset provides medical terminology in isiXhosa, addressing a critical gap in resources for this domain. This specialised dataset enables the application of lexicon-based approaches like DALI [18] and DiPMT [22] that specifically target terminology challenges [22]. This dataset opens new possibilities for combining established methods to address both low-resource challenges and domain specialisation, with the potential of creating more accurate medical translation systems for isiXhosa speakers.

7 CONCLUSIONS

In this literature review, several interconnected topics were explored in the field of machine translation with a specific focus on isiXhosa in the medical domain. The literature review began by examining the evolution of MT architectures, from early encoder-decoder models to attention mechanisms that revolutionised translation quality. The Transformer architecture that now forms the backbone of state-of-the-art MT systems was analysed and compared with decoder-only models like large language models. The review then investigated the unique challenges of isiXhosa translation, including data scarcity, morphological complexity, and dialectal differences that make it particularly difficult for MT systems.

Several promising approaches for low-resource language translation were identified. Multilingual NMT has shown significant benefits by leveraging related languages to improve isiXhosa translation performance, with studies demonstrating substantial BLEU score improvements. Transfer learning from related Nguni languages has proven especially effective, allowing models to capitalise on linguistic similarities. Data augmentation strategies, particularly back-translation, have helped address the fundamental issue of limited parallel data.

Domain adaptation represents another critical dimension when considering specialized fields like healthcare. The review highlighted both data-centric and model-centric approaches, with techniques such as fine-tuning, synthetic data generation, and lexicon-based methods showing particular promise for medical terminology. Lexicon-based approaches like DALI and DiPMT have demonstrated effectiveness in handling specialised terminology, which is crucial for medical translation where accuracy is crucial.

The Blocker et al. dataset represents a significant contribution to the field, providing authentic parallel data in a domain where resources have been extremely limited. This dataset addresses the dual challenges of low-resource language translation and domain-specific adaptation. By applying established methods like back-translation and lexicon-based approaches to this new domain-specific dataset, researchers can work toward improving the accuracy of isiXhosa medical translation.

REFERENCES

[1] Huang, J., Li, J., and Xue, H. (2020). *A Comparative Study of RBMT, SMT, and NMT for Low-Resource Domain Adaptation.* MDPI Information, 11(5), 259.

[2] Bahdanau, D., Cho, K., and Bengio, Y. (2015). *Neural Machine Translation by Jointly Learning to Align and Translate.* https://arxiv.org/abs/1409.0473.

[3] Sennrich, R., Haddow, B., and Birch, A. (2016). *Neural Machine Translation of Rare Words with Subword Units.* In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725.

[4] Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). *No Language Left Behind: Scaling Human-Centered Machine Translation.* In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5246–5277.

[5] Restack, A. (2023). *Quantitative Analysis of English–isiXhosa Parallel Datasets.* Journal of Low-Resource MT, 12(2): 1–12.

[6] Adelani, D. I., Alam, M. I., Anastasopoulos, A., *et al.* (2022). *Findings of the WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages.* In *Proceedings of the Seventh Conference on Machine Translation (WMT 2022)*, 773–800.

[7] Wang, L., et al. (2023). *Document-level Machine Translation with Large Language Models.* https://arxiv.org/abs/2304.02210.

[8] Niu, X., Qu, Y., and Neubig, G. (2023). *Evaluating ChatGPT as a Low-Resource MT Model.* In *Proceedings of the 2023 ACL.*

[9] Levin, M. (2006). *Language and Health in South Africa: The Role of African Languages in Healthcare.* African Journal of Health, 22(3).

[10] Chu, C., and Wang, R. (2020). *A Survey of Domain Adaptation for Machine Translation.* Journal of Information Processing, 28, 413–426.

[11] Koehn, P., and Knowles, R. (2017). *Six Challenges for Neural Machine Translation.* In *Proceedings of the First Workshop on Neural Machine Translation*, 28–39.

[12] Saunders, D. (2021). *Survey of Domain Adaptation in Neural Machine Translation.* https://arxiv.org/pdf/2104.06951.

[13] Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). *Understanding Back-Translation at Scale.* In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 489–500.

[14] Sennrich, R., Haddow, B., and Birch, A. (2017). *Neural Machine Translation of Rare Words with Subword Units: Revisiting Segmentation Strategies.* In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 66–75.

[15] Jin, Z., Feng, W., and Li, M. (2020). *Domain Adaptation via Back-Translation for Neural Machine Translation.* In *Proceedings of the 2020 Conference on Computational Natural Language Learning (CoNLL)*, 54–64.

[16] Chinea-Ríos, M., Martínez, L., Toma, I. et al. (2017). *Self-Learning Approaches for Low-Resource Neural Machine Translation.* In *Proceedings of the International Conference on Natural Language and Speech Processing*, 90–99.

[17] Zhang, J., and Zong, C. (2016b). *Bridging Source and Target Embeddings for Cross-Domain Neural Machine Translation.* In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1993–2003.

[18] Hu, J., Neubig, G., Oda, Y., et al. (2019). *Domain Adaptation by Lexicon Induction for Neural Machine Translation.* In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2985–2995.

[19] Artetxe, M., Labaka, G., and Agirre, E. (2019). *An Effective Approach to Unsupervised Machine Translation.* In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 194–203.

[20] Zhang, S., Liu, Y., and Zhou, M. (2020). *Dictionary-based Data Augmentation for Cross-Domain Neural Machine Translation.* In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7987–7998.

[21] Rihan, F., et al. (2023). *Integrating LLM-Generated Synthetic Data for Domain Terminology in Low-Resource MT.* In *Proceedings of the Eighth Conference on Machine Translation.*

[22] Zheng, X., et al. (2023). *Dictionary-Informed Prompt-Based Machine Translation.* https://arxiv.org/pdf/2302.07856.

[23] Robinson, N. R., Ogayo, P., Mortensen, D. R., and Neubig, G. (2023). *ChatGPT MT: Competitive for High- (but not Low-) Resource Languages.* arXiv:2309.07423 [cs.CL].

[24] Cho, K., van Merri"enboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.* In *Proceedings of the 2014 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734.

[25] Luong, M.-T., Pham, H., and Manning, C. D. (2015). *Effective Approaches to Attention-based Neural Machine Translation.* In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421.

[26] Popel, M., and Bojar, O. (2018). *Training Tips for the Transformer Model.* The Prague Bulletin of Mathematical Linguistics, 110(1), 43–70.

[27] Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., et al. (2021). *Beyond English-Centric Multilingual Machine Translation.* Journal of Machine Learning Research, 22(107), 1–48.

[28] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). *Sequence to Sequence Learning with Neural Networks.* In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3104–3112.

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). *Attention is All You Need.* In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

[30] E. Nyoni and B. A. Bassett. (2021). *Low-Resource Neural Machine Translation for Southern African Languages.* https://arxiv.org/abs/2104.00366.

[31] C. C. Emezue, B. F. P. Dossou, and the Masakhane Team. (2021). *MMTAfrica: Multilingual Machine Translation for African Languages.* In *Proceedings of the Sixth Conference on Machine Translation (WMT 2021)*, 398–411.

[32] C. C. Emezue, B. F. P. Dossou, and the Masakhane Team. (2022). *Large-Scale Many-to-Many Multilingual Translation for African Languages.* In *Proceedings of the Seventh Conference on Machine Translation (WMT 2022).*

[33] K. N. Elmadani, F. Meyer, and J. Buys. 2022. University of Cape Town's WMT22 system: Multilingual machine translation for southern African languages. *arXiv preprint* arXiv:2210.11757.

[34] Blocker, A., Meyer, F., Biyabani, A., Mwangama, J., Datay, M. I., and Malila, B. (2025). *Benchmarking IsiXhosa Automatic Speech Recognition and Machine Translation for Digital Health Provision.* Proceedings of the Workshop on Patient-oriented Language Processing, NAACL 2025.

[35] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). *Attention is All You Need.* Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.

[36] Levin, M. (2006). *Language and Health in South Africa: The Role of African Languages in Healthcare.* African Journal of Health, 22(3): 211–219.

[37] Wang, Y., Chen, Z., and Liu, K. (2022). *Domain Adaptation in Neural Machine Translation: A Survey.* ACM Computing Surveys, 55(3):1–34.

[38] Hutchins, W.J., and Somers, H.L. (1992). *An Introduction to Machine Translation.* London: Academic Press. Chapter 4: Rule-Based Machine Translation Systems, 65–96.

[39] Bahdanau, D., Cho, K., and Bengio, Y. (2015). *Neural Machine Translation by Jointly Learning to Align and Translate.* In *Proceedings of the International Conference on Learning Representations (ICLR).*

[40] Goyal, N., Gao, C., Chaudhary, V., Chen, P., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). *The Flores-101 evaluation benchmark for low-resource and multilingual machine translation.* Transactions of the Association for Computational Linguistics, 10:522–538.

[41] Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). *A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures.* Neural Computation, 31(7):1235–1270.