Literature Review on Domain-Specific and isiXhosa Machine Translation

Elijah Sherman (SHRELI006)

ABSTRACT

Machine translation (MT) has made significant advancements, particularly with the development of Neural Machine Translation (NMT) and the transformer architecture. However, these improvements are not evenly distributed across all languages. Low-resource languages such as isiXhosa continue to experience challenges in MT due to the limited availability of parallel corpora. This literature review explores recent developments in NMT, focusing on their applicability to isiXhosa and domain-specific translation, particularly in the medical field. This review examines various MT approaches, including statistical and neural methods, highlighting the advantages of NMT in producing contextually accurate translations. It also discusses key architectures such as encoder-decoder models, transformers, and decoder-only models used in large language models (LLMs). Additionally, it addresses the unique challenges of MT for isiXhosa, considering the scarcity of training data and the effectiveness of techniques like transfer learning, zero-shot learning, and multilingual modelling. A significant focus is placed on domain adaptation for MT, particularly in the medical field, where accurate translations can bridge communication gaps between doctors and isiXhosa-speaking patients. Various strategies for improving domain-specific translation are explored, including domain-adapted fine-tuning, data augmentation, and prompt engineering for LLMs. This review provides insights into how these approaches can enhance English-isiXhosa translation, contributing to more effective and accessible MT solutions for critical applications.

1 INTRODUCTION

Machine translation (MT) has seen significant improvements recently with research being done on different approaches and architectures. This includes Neural Machine Translation (NMT) which utilises deep neural networks to learn the context of the sentence to generate a more semantically correct translation. The introduction of the transformer architecture [22] further improved NMT models by relying purely on an attention mechanism, allowing the model to make more accurate translations specific to the context of the sentence.

These advancements in MT are not universal as MT is still not as effective on low-resource languages (such as isiXhosa and many other African languages) as it is on high-resource languages (such as English, French, German etc.). This problem is prevalent in South Africa as most of its 12 official languages are low-resource languages. One of these languages is isiXhosa which is one of the most widely spoken languages in the country, meaning there are many people who are not able to utilise MT since it is not as effective when translating in their home language.

This inaccessibility has lead to further inaccessibility of isiXhosa speakers to better medical care. This problem arises because many doctors in South Africa only speak English fluently, which creates a language barrier between them and many of their patients. This barrier hinders communication and can have a negative impact on patient care. This highlights the need for effective MT for English-isiXhosa in the medical domain. An advancement in this field would allow for effective doctor-patient communication, giving doctors a better understanding of the patient's symptoms and giving the patient a better understanding of the doctors explanation.

Building a MT model that can effectively translate between English and isiXhosa in the medical domain poses significant challenges since isiXhosa is a low-resource language and training models for a specific domain can be complex. The complexity of the medical domain compounds this challenge as it has many complicated terms that may not have a direct translation from English to isiXhosa. There are various techniques to alleviate this challenge such as domain-specific prompting when using LLMs for MT and training the model on synthetic data which allows it to gather in-domain knowledge to make more accurate predictions on words that are out of vocabulary.

There has already been research done on this topic (English-isiXhosa MT in the medical domain). Recently Blocker et al. (2025) gather new data on English and isiXhosa medical texts using automatic speech recognition (ASR) and evaluates the performance of various MT models when translating using these texts. This will be essential in our project which aims to improve machine translation in the medical domain for isiXhosa. Our project will achieve this by exploring various options for adapting existing English-isiXhosa MT models to the medical domain.

2 NEURAL MACHINE TRANSLATION

Machine translation (MT) is the translation of human languages from one to another by a computer. Traditionally, this is done by having many sub-components that are separately trained [18], however, Neural Machine Translation (NMT) takes a different approach and aims to build and train a single neural network that performs the entire translation process [5]. Both Statistical Machine Translation (SMT) and NMT are trained using parallel corpora with sentences from one language mapped to another.

2.1 Statistical Machine Translation

Statistical machine translation models analyse smaller pieces of text such as words or phrases rather than the entire sentence. It looks at each piece and uses statistical patterns to identify how often the word or phrase occurs with particular words in the parallel corpora [5]. Using these patterns they determine the most likely translation for the word or phrase and after translating each segment of the source sentence, it combines them and outputs the whole translated sentence [14].

2.2 Neural Machine Translation

Neural machine translation is an improvement to SMT models as it uses neural networks and deep learning on patterns in the data instead of just translating individual words or phrases based on how frequently they co-occur in the parallel corpora [23]. NMT models translate the entire sentence in an end-to-end fashion taking into account the context of the sentence when translating each word to increase the probability of an accurate translation [23].

Instead of only mapping whole words to each other, tokenisation is used to match sub-words in the sentence [12] which makes translations more accurate for certain languages where there is not a one-to-one mapping of a word from the source language to the target language. Tokenisation is done by running a tokenisation algorithm on the training corpus. There are various algorithms that can be used. For example, the wordpiece algorithm is a powerful tokenisation algorithm that chooses tokens based on which one increases the probability of the language model tokenisation the most [12].

2.2.1 Encoder-Decoder Models. Sequence-to-sequence models are common for NMT and they utilise the encoder-decoder framework [23]. First the encoder takes the source sentence as input and converts each input symbol into a vector. This list of vectors is passed to the decoder which uses them to produce one symbol at a time [23]. These vectors are created using word embedding that gives the words more meaning that can be used to make the semantics of the output sentence more accurate [3].

2.2.2 Decoder Only Models. Large Language Models (LLM) can be used for MT and are often implemented as decoder only models [15]. LLMs work by predicting the next token given a certain input. In the context of MT they do not process the input sentence but rather, given a prompt such as "{The input sentence} translate this sentence into isiXhosa:", it will predict the next tokens based on the prompt giving the translation of the input sentence [8]. This has been shown to be quite effective in multilingual MT ,however , LLMs do not perform as effectively with low-resource languages [15].

2.3 Neural Architectures

There are two main architectures that we will talk about: Long Short-Term Memory (LSTM) [10] and Transformers [22].

2.3.1 Long Short-Term Memory. LSTMs are a type of Recurrent Neural Network (RNN) that use an internal state to store semantic information to make better predictions in the next step.

A RNN is a neural network that receives two inputs, the first being the regular input and the second being the output from the previous step known as the hidden state. The second input is given to the network because each step depends on the previous step and requires memory of it to make a suitable prediction. One of the main problems with RNNs is the vanishing gradient problem which refers to the gradients shrinking during the training of the neural network. This problem occurs when the gradients becoming increasingly small during backwards propagation which makes it difficult to successfully update the weights in the network [4].

LSTMs aim to get rid of this problem by using an internal state as well as the hidden states to make better predictions [13]. This internal state consists of three gates: a forget gate, an input gate and an output gate. Each gate has a value between zero and one that determines what state information can be forgotten (forget gate), what information should be added to the state (input gate) and what information should be output in this instance (output gate). The closer to zero the value for particular information is, less of that information will get through the gate [13]. These gates will help the network to make better predictions at each step.

2.3.2 Transformers. The transformer architecture relies entirely on an attention mechanism to make predictions [22] and is used to replace RNNs. Attention is a process that connects queries and a set of key-value pairs to an output [22].

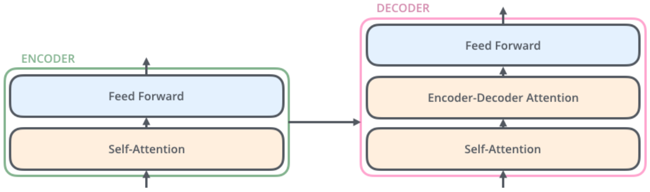

Figure 1: Sub-Layers within the Transformer Encoder-Decoder Architecture

In a transformer, an encoder has a self-attention layer and a feed forward neural network layer. A decoder has the same layers with an additional layer in between them called the encoder-decoder attention layer. These layers can be seen in figure 1. The self-attention layer allows the transformer to look at other words in the sentence while it encodes a word [2]. The encoder-decoder attention layer allows the decoder to concentrate on the most relevant pieces of the rest of the sentence to generate a more accurate output [2].

Transformers use a type of attention called multi-head attention. This type of attention allows the transformer to use the attention function in parallel, giving it access to more information simultaneously at each time step [22]. For example, in the encoder-decoder attention layer, queries come from the previous decoder layer and the key-value pairs come from the output of the encoder. This collection of information allows the decoder to be aware of all positions in the input sequence at each time step [22].

3 ISIXHOSA TRANSLATION

isiXhosa is widely spoken in South Africa, making it essential for medical professionals to understand, as many of their patients are native speakers.

3.1 MT for Low-Resource Languages

An essential part of training language models for MT is the corpora of data that are used to train them. For some languages such as English, French and German there are many data sets available for training purposes as these languages are widely spoken around the world. These are known as high-resource languages, and they are contrasted to low-resource languages which are not spoken by as many people and therefore do not have as much data to train language models on. Many African languages fall into this category meaning MT for these languages is not very advanced [17].

Since NMT is preferred over traditional MT it is often used for these low-resource languages, however, this poses a problem as developing an NMT system is data-demanding [17]. Nyoni and Basset (2021) explore various techniques that can be used specifically for low-resource language models, namely: transfer learning, zero-shot learning and multilingual modelling. After their experiments they found that multilingual modelling performed the best on their dataset but transfer learning is a successful technique for training models on low-resource languages that are similar to each other[17]. For example, using their English-isiXhosa model to train an English-isiZulu model proved to be very effective with an improvement of 9.9 in the BLEU score.

3.2 isiXhosa Data for NMT

Since isiXhosa is a low-resource language, there have been attempts to increase the data that is available and improve the quality of MT. Adelani et al. (2022) present the findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages which shows how a variety of models perform when translating between a set of African languages and English and French[1]. The best performing model from this study was a large transformer model that grouped target languages together to fine-tune separate models for each group [1]. The overall findings of this study is that data is still very important for the quality of translations in MT but progress is being made for MT in low-resource African languages by increasing the data that is available. To evaluate these translations the FLORES-101 benchmark [9] was used as it recently expanded to include more African languages. The advantages of using this benchmark is that it allows for many-to-many evaluation and covers a large number of languages across a variety of domains [9].

4 DOMAIN-SPECIFIC TRANSLATION

When translating from one language to another, the context is very important to ensure an accurate translation. When the source text is part of a specific domain, translation is often more complicated since it could contain rare words that the MT model is unaware of. Another possibility is certain words could have different meanings in the context of this specific domain compared to when they are used in another domain.

4.1 Challenges in Domain-Specific Translation

When training a MT model for a specific domain such as mathematics, computer science, medicine etc., it performs better in that domain but worse for any other domain [11]. This is because a domain has specific terms associated with it and if the training data did not contain those terms it will be unable to translate them correctly.

The medical domain is particularly complex and involves many concepts that are difficult to understand. This makes translating from one language to another in this domain hard because there are lengthy words and terms that could be rare and might not have a direct translation in the target language.

Because of the sensitive nature of the domain, translations need to be precise and accurate otherwise it could lead to misunderstandings that can compromise the patient's well-being [20]. For example, in the experiment done by Pang et al. (2025) their medical translation system translated "Tatort" to "accident" instead of "crime". Errors such as this have the potential to have a negative impact on a patient which is why it is difficult to create an effective MT model for the medical domain.

4.2 Training Models on Synthetic Data

Hu et al. (2019) propose a solution to this problem called domain adaptation by lexicon induction (DALI). This approach uses large amounts of synthetic data to create a pseudo-parallel corpus of in-domain data. Their experiments found that DALI works well to fine-tune a NMT model to effectively translate out of domain words [11]. The problem with this approach is that it requires a large corpus of data to construct the pseudo-parallel corpus and for low-resource languages such as isiXhosa this is not easily accessible.

Moslem et al. (2022) propose a different solution which uses language models (LM) to enhance data for a different domain that the MT model is not familiar with. This method generates a large number of sentences which simulates either a small bilingual dataset which allows the MT model to learn domain-specific patterns or a monolingual text which LMs can use to generate synthetic translations [16]. They found in their experiment that these models show significant improvements when being automatically evaluated using the BLEU scoring system [21] and when being evaluated by humans.

Comparing the average BLEU results of these two approaches it is clear the approach taken by Moslem et al. (2022) is better as it shows more significant improvements when translating out of domain text.

4.3 Domain-Specific Prompting of LLMs

While LLMs are not specifically designed for MT they can still be quite effective for this purpose [8]. Ghazvininejad, Gonen & Zettlemoyer (2023) introduce the idea of Dictionary Based Prompting for Machine Translation (DiPMT) which shows improvements in MT using LLMs for both low-resource languages and out of domain words. The method behind DiPMT is to give one or more key translations for the sentence in the prompt to the LLM [8]. For example:

> "Translate the following sentence to English: "*Pada dasarnya, hal tersebut terbagi ke dalam dua kategori: Anda bekerja sambil mengadakan perjalanan atau mencoba mencoba atau membatasi pengeluaran Anda. Artikel ini berfokus pada hal yang terakhir.*"
> In this context, the word "sambil" means "while"; the word "membatasi" means "limiting", "restrict", "limit".
> Output: The full translation to English is: Basically, they fall into two categories: Either work while you travel or try and limit your expenses. This article is focused on the latter." [8]

5 DISCUSSION

NMT has made significant progress over the last decade[5][7] as more research has been done on it. NMT has been proven to be more accurate than SMT as it uses deep neural networks to leverage contextual information on the sentence. This information improves sentence accuracy and semantic correctness, which SMT models typically handle less effectively. The transformer architecture [22] has improved NMT further as it relies mainly on an attention mechanism to gather contextual information in the sentence. More specifically, they use multi-headed attention which allows it to use more information at the same time when translating a sentence.

While isiXhosa MT has seen some improvement recently, it is far behind MT for high-resource languages such as French and English. Synthetic data has been shown to improve performance of models trained on low-resource languages such as isiXhosa. There are various approaches to create synthetic data that can be used to train MT models but the LM approach [16] has the strongest performance. This approach has been shown to effectively utilise the medical domain data that is available for isiXhosa and find patterns within the data to make better predictions for out of vocabulary words.

This approach can be combined with DiPMT [8] (where certain terminology is defined in the prompt to the LLM) for domain-specific translation to further improve our NMT model's ability to translate within the medical domain. For example if a doctor has diagnosed a patient with arthritis they give the following prompt to the LLM:

> Translate the following sentence into isiXhosa: "Arthritis is a condition that causes inflammation and stiffness in the joints, leading to pain and reduced movement." In this context, the word "arthritis" means "isifo samathambo".

These techniques are potentially relevant to the recent work done by Blocker et al. (2025) who created a dataset specifically for isiXhosa in the medical domain. This data was created using transcribed medical consultations from a pre-existing dataset. These transcriptions were then read by South African actors in both English and isiXhosa and the audio was recorded to be transcribed by automatic speech recognition (ASR) models. Some of the models used were open-source and some were commercial models, with one of the open-source models, Whisper [19], performing the best. These newly transcribed texts were then put through various MT models to assess how well they perform translations on these texts. These models also varied between open-source and commercial models with some models being dedicated MT models and some being LLMs. The performance of each of the MT models varied, with ChatGPT performing the best for isiXhosa to English translations and Google Cloud Translate performing best for English to isiXhosa [6].

This dataset is important as it allows for further training to be done for MT models however it also presents a challenge as it is both low-resource and domain-specific. The domain adaptation techniques we discussed will be helpful to improve the performance on this dataset, allowing for the further development of MT for isiXhosa in the medical domain.

6 CONCLUSIONS

In this literature review, we discussed topics related to NMT, isiXhosa translation, and domain-specific translation. We began by comparing SMT and NMT, introducing state-of-the-art NMT techniques with a focus on the transformer architecture, which utilizes an attention mechanism. Next, we assessed the current state of MT for isiXhosa, highlighting how most MT models perform poorly for low-resource languages. We reviewed the performance of various techniques for improving low-resource language models and found that multilingual modeling yields the strongest results. We then examined the challenges posed by domain-specific translation and explored two approaches to address them: generating synthetic data and domain-specific prompting of LLMs. Both techniques have their merits and enhance the feasibility of effective MT within a specific domain.

The performance of existing MT models on the dataset compiled by Blocker et al. (2025) underscores the need for improved MT models tailored to isiXhosa in the medical domain. These results highlight the potential of domain adaptation, as the discussed techniques can contribute to advancing MT for isiXhosa in medical contexts.

REFERENCES

[1] Adelani, D. I., Alam, M. M. I., Anastasopoulos, A., Bhagia, A., Costa-jussà, M. R., Dodge, J., Faisal, F., Federmann, C., Fedorova, N., Guzmán, F., Koshelev, S., Maillard, J., Marivate, V., Mbuya, J., Mourachko, A., Saleem, S., Schwenk, H., and Wenzek, G. 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)* (Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022), P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, Eds., Association for Computational Linguistics, pp. 773–800.

[2] Alammar, J. The illustrated transformer, 2018.

[3] Alammar, J. Visualizing neural machine translation mechanisms of seq2seq models with attention, 2018.

[4] Amanatullah. Vanishing gradient problem in deep learning: Understanding, intuition, and solutions, Jun 2023.

[5] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate, 2016.

[6] Blocker, A., Meyer, F., Biyabani, A., Mwangama, J., Datay, M. I., and Malila, B. Benchmarking isixhosa automatic speech recognition and machine translation for digital health provision. In *Proceedings of the Workshop on Patient-oriented Language Processing* (2025).

[7] Buscaldi, D., and Rosso, P. How good is nllb-200 for low-resource languages? a study on genoese. In *CLiC-it 2023: 9th Italian Conference on Computational Linguistics* (2023).

[8] Ghazvininejad, M., Gonen, H., and Zettlemoyer, L. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856* (2023).

[9] Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics 10* (05 2022), 522–538.

[10] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[11] Hu, J., Xia, M., Neubig, G., and Carbonell, J. Domain adaptation of neural machine translation by lexicon induction, 2019.

[12] Jurafsky, D., and Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. Stanford University, 2025. Online manuscript released January 12, 2025.

[13] Keen, M. What is lstm (long short term memory)?, Nov 2021.

[14] Lopez, A. Statistical machine translation. *ACM Comput. Surv. 40*, 3 (Aug. 2008).

[15] M., A. P., M, S. R., and Christopher, O. Machine translation with large language models: Decoder only vs. encoder-decoder, 2024.

[16] Moslem, Y., Haque, R., Kelleher, J., and Way, A. Domain-specific text generation for machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (Orlando, USA, Sept. 2022), K. Duh and F. Guzmán, Eds., Association for Machine Translation in the Americas, pp. 14–30.

[17] Nyoni, E., and Bassett, B. A. Low-resource neural machine translation for southern african languages, 2021.

[18] Omniscien Technologies. What is neural machine translation?, n.d. Accessed: 2025-03-10.

[19] OpenAI. Whisper on hugging face spaces. https://huggingface.co/spaces/openai/whisper, 2025.

[20] Pang, J., Ye, F., Wong, D. F., Yu, D., Shi, S., Tu, Z., and Wang, L. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics 13* (01 2025), 73–95.

[21] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (2002), pp. 311–318.

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[23] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.