Lexicon Induction for isiXhosa Medical Translation

Elijah Sherman
shreli006@myuct.ac.za
Department of Computer Science, University of Cape Town
Cape Town, South Africa

ABSTRACT

Machine Translation (MT) for low-resource languages such as isiXhosa is challenging in specialised domains such as medicine. Domain adaptation by lexicon induction (DALI) is a method for creating synthetic parallel corpora for low resource languages in a specific domain. This is done by word-for-word back-translating a monolingual corpus of data using a bilingual lexicon. DALI achieved performance gains in adapting MT models to specific domains in the initial study introducing the technique. However, it has not been tested in a truly low-resource setting such as isiXhosa medical. This project tests the efficacy of DALI generated synthetic data for fine-tuning models for English-isiXhosa MT in the medical domain. The results of the experiments done indicate that DALI does not improve medical domain translation for isiXhosa. For Eng→Xho translation the baseline model outperforms all the fine-tuned models, while for Xho→Eng translation the fine-tuned models marginally outperform the baseline model. These models were also compared to others trained by team-mates. The models fine-tuned using back-translated synthetic data significantly outperformed the DALI models in both translation directions. This could be because DALI is not suitable for truly low-resource languages such as isiXhosa and requires a higher quality lexicon, with a larger vocabulary to perform better. The health term error rate was calculated across various categories which shows the lexicon may not be of a high enough quality, as the baseline typically outperformed the fine-tuned models in all categories.

1 INTRODUCTION

Machine translation (MT) is the process of translating text from one language to another using a computer. Recently there have been major advancements in the MT field with the introduction of Neural Machine Translation (NMT) which trains a single neural network to perform the entire translation process [4].

There are however significant problems in certain areas of MT which require more research. Two of these problem areas are MT for low-resource languages and domain specific MT. Low-resource languages are those that have very little data, specifically parallel corpora, available for training MT models. This means that MT done for these languages is often of a lower quality since deep learning is completely data-driven. Domain specific MT refers to models being trained or fine-tuned for a specific domain to improve their performance when translating in that domain. General domain MT models often experience a drop in performance when translating in a specific domain or if a model is trained on one domain it often performs worse in another [17]. This is because the training data used lacks domain-specific terminology, meaning the model does not know how to translate these terms. This problem is exacerbated by a lack of in-domain data.

IsiXhosa is a low-resource language meaning that current MT models do not perform as well when translating isiXhosa as they do when translating high-resource languages [1]. This is a crucial problem as isiXhosa is one of the most widely spoken home languages in South Africa [29] and without effective MT models many South Africans are excluded from using this helpful technology. Additionally, many doctors in South Africa only speak English while many of their patients may only speak isiXhosa. This poses a serious problem as effective communication between doctors and their patients is crucial to ensure the necessary treatment is given and a human translator may not always be available. This highlights the need for effective MT to assist doctors in understanding their patients and vice-versa.

This project attempts to further the research done in this area by exploring domain adaptation by lexicon induction (DALI) [14] as a method for generating synthetic data. This method will be used for training MT models for English ↔ isiXhosa translation in the medical domain. DALI generates a bilingual lexicon which is used to word-for-word back-translate a large monolingual corpus into the source language, thereby creating a pseudo-parallel corpus of training data. This approach is particularly well-suited for isiXhosa medical MT as it does not require large amounts of in-domain parallel corpora to train the models.

In this project, the pseudo-parallel corpus was used to fine-tune the NLLB-200 model [10] using various hyperparameters to find the optimal set for both Eng→Xho and Xho→Eng translation. These models were initially evaluated on two datasets, Flores-200 [21] and MeMaT medical [16], to find the best ones, and produced BLEU [25] scores between 5.98-9.63 and 21.82-26.31 for Eng→Xho and Xho→Eng translation respectively.

The main evaluation set used was the Blocker evaluation dataset [6] which is a parallel corpus of mock doctor-patient consultations. This dataset was used because it is in-domain and was human translated making it very reliable for evaluation. The best performing models were evaluated on the Blocker dataset with the baseline NLLB-200 [10] outperforming the Eng→Xho translation models and both of the Xho→Eng translation models outperforming the baseline.

These results were also compared to alternative approaches to solving this problem done by team members. This includes two baseline models and an alternative approach to domain adaptation. The two baselines are only trained on general domain data to test how well they would perform on in-domain evaluation. The first baseline is NLLB-200 fine-tuned on the Eng-Xho subset of the WMT22 dataset [9] and the second is a model trained from scratch using that same dataset. The alternative approach is to fine-tune NLLB-200 on synthetic data generated back-translation.

The results for Eng→Xho translation show the baseline NLLB model outperforms the fine-tuned models with a difference in BLEU

score of approximately 1. This is likely due to the small size of the training data which may have caused overfitting. The results for Xho→Eng translation show the fine-tuned model outperforms the baseline by 1.11 BLEU. This improvement shows that the model has been adapted to the medical domain.

However, these results are contradicted by the health term error rates that were calculated across various categories. These results show that the baseline typically had the lowest error rate over most of the categories, meaning the general domain model was the most accurate when translating medical terminology. This is likely due to overfitting and catastrophic forgetting caused by the small training set.

2 RELATED WORKS

2.1 Neural Machine Translation

Statistical machine translation (SMT) was the dominant approach for MT before the introduction of neural networks. SMT is done by analysing co-occurrence patterns of words or phrases in large parallel corpora [19]. This approach is effective, however, it relies on a series of separately trained sub-components [24]. Because of this, it often struggles to capture the context and natural tone of human language as it generates translations [5].

Neural machine translation (NMT) is an improvement over SMT as it uses a neural network to perform the entire translation process [4]. Unlike the co-occurrence analysis done in SMT, NMT models leverage deep learning to understand the context of the input sentence to generate a more accurate and natural sounding translation [31]. This is aided by the use of tokenisation which is used to break words down into sub-words [15]. Tokenisation is especially useful for highly agglutinative languages where complex words are formed by many sub-words.

The transformer architecture for NMT introduced by Vaswani et al. (2017) is the state-of-the-art architecture used for NMT. Transformers use the encoder-decoder model architecture and replace the recurrent layers of previous architectures with a self-attention mechanism [30]. This self-attention mechanism, called multi-head attention, allows the model to process more information simultaneously, enabling it to use the context of the sentence to generate more accurate translations. The initial experiments done using transformers show that these models can be trained faster and gain performance improvements over previous architectures.

2.2 IsiXhosa Translation

The state of MT has greatly advanced over the last decade with the advent of NMT. These advancements are easier to make for languages such as English, French and Spanish as they are some of the most widely spoken languages in the world. These are known as high-resource languages because of the large-scale of publicly available datasets that are crucial for training high-performing models. Conversely, many African languages are known as low-resource languages because of a clear lack of training data available. Because of this, MT models do not perform well when translating these languages and will give lower quality or incorrect translations [22].

IsiXhosa is one of these low-resource languages, creating a need for more research to be done to improve the state of MT for it. Research has been done on using various techniques to maximise the utility of available data to improve MT models. Nyoni and Basset (2021) explored three of these techniques: transfer learning, zero-shot learning and multilingual modelling. These techniques were tested on improving MT models for isiZulu, isiXhosa and Shona. They found that multilingual learning performed the best but all three of the techniques resulted in improved BLEU scores over their baseline model [22]. What was especially interesting is they found transfer learning to be particularly effective for training low-resource models for closely related languages, namely training their English-isiZulu model using their English-isiXhosa model showed a significant increase in the BLEU score.

These techniques are what allow multilingual NMT models to perform better than standard bilingual models, even in low-resource settings. Multilingual MT models use a single neural network to translate between many source and target languages [2]. One such model is Meta's No Language Left Behind 200 (NLLB-200) model, which can translate between 200 languages including isiXhosa [10]. Transfer learning is the key to why these models are so effective. When training on a variety of languages, the model may be able to use knowledge from high-resource languages to improve the quality of translations for low-resource languages. An additional benefit of this is zero-shot translation. This is where a model that can translate from English → French and English → isiXhosa can often translate from French → isiXhosa without ever being trained to do so.

Another key contribution to isiXhosa MT research is the Workshop on Machine Translation (WMT) in 2022 which included evaluation for African languages [1]. This workshop made datasets for low-resource languages openly available for other researchers to use which will allow for further exploration into low-resource MT techniques.

2.3 Domain Specific Translation

MT models struggle when translating domain specific sentences as they often do not know how to translate specialised terminology [17]. This creates a need for specially trained models for these domains, however this can be challenging due to a lack of parallel corpora available.

Synthetic data can be generated to compensate for the lack of real data. There are many methods for doing so but back-translation is the most common. Back-translation involves using an MT model to translate text from the target language into the source language, thereby giving a pseudo-parallel corpus of training data [8]. There are other, more technical approaches to this such as DALI which will be discussed further in section 3. It has been shown that synthetic data has the potential to be just as effective as real data, if it is of a high enough quality.

Another avenue of approaches to generating synthetic data is to use language models (LM). Moslem et al. (2022) investigate a method for generating synthetic domain-specific data using LMs. Their technique feeds a pre-trained LM a large monolingual corpus of in-domain data in the source language and uses it to generate additional data [20]. These additional sentences are then translated into the target language by a general-domain MT model. This results in a pseudo-parallel corpus which can be used to fine-tune

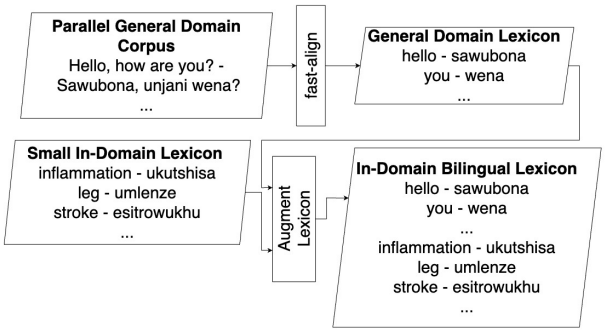

Figure 1: Medical Domain Lexicon Creation

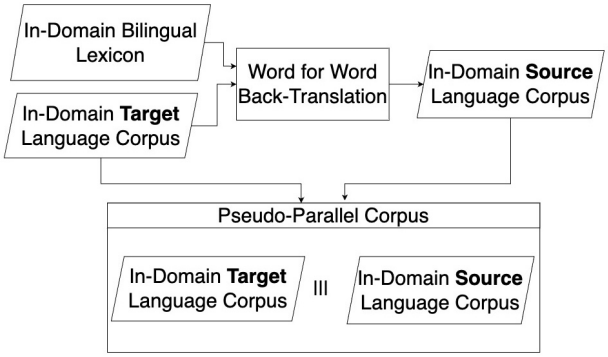

Figure 2: Pseudo-Parallel Corpus Generation

models to the target domain. This was shown to be effective as the synthetic data was more lexically diverse, giving the MT model a more comprehensive vocabulary. However, this technique is flawed as it relies on a general-domain MT model to translate in-domain vocabulary which likely results in incorrect translations. These incorrect translations will be carried over to the models trained on this data, hindering its performance.

Another approach was proposed by Ghazvininejad, Gonen & Zettlemoyer (2023) called Dictionary Based Prompting for Machine Translation (DiPMT) [12]. DiPMT uses large language models (LLM) which are capable of MT. In the translation prompt to the model certain key word translations will be given. This has been shown to improve the quality of the translations given. However, this approach is not ideal as it requires key translations to be given every time the model is prompted in order to be effective.

3 OUR APPROACH: DALI

This project is based on the work done by Hu et al. (2019) who explored adapting NMT models to specific domains using an unsupervised, data-based approach called domain adaptation by lexicon induction (DALI) [14].

NMT models that have not been adapted to any particular domain do not perform well when evaluated on domain specific data due to their inability to translate words that are unknown, or out of vocabulary (OOV). Similarly models that have been adapted to a specific domain do not perform well when evaluated on data in a vastly different domain. DALI aims to solve this issue by investigating domain adaptation methods that correctly translate words that are OOV. DALI adapts NMT models to a domain by generating a pseudo-parallel corpus of training data using a monolingual corpus and a bilingual lexicon. This is done in two main steps: lexicon generation and synthetic data generation.

3.1 Lexicon Generation

First the lexicon is generated by taking a general domain parallel corpus and running a word-alignment program such as GIZA++ [23] or fast-align [7]. This gives a general domain lexicon. This lexicon is adapted to the domain by adding a relatively small dictionary containing key in-domain words, resulting in a large in-domain bilingual lexicon. This process is shown in Figure 1.

This lexicon is well-suited for low-resource languages as it can be automatically created from existing corpora, generating a large list of word pairs. The seed lexicon generated from the general domain parallel corpus contains many basic words, giving it a vast vocabulary. The explicit introduction of the in-domain terminology to this lexicon increases the vocabulary to include these words without excluding the basic words needed for most translations. The lexicon is also a good fit for low-resource languages because it bypasses the need for a pre-trained MT model to perform back-translation.

However, the lexicon is often filled with noise created by mistakes made by the word-alignment tool used. These mistakes are then spread to the synthetic dataset, causing it to be of a lower quality. IsiXhosa is a highly agglutinative language, meaning that a single word in isiXhosa could translate to an entire sentence in English. This causes more mistakes to be made by word-alignment tools when used on highly agglutinative languages. Another critical short-fall of these tools is their inability to capture the context of the sentence they are extracting words from. This means that homonyms used in the general domain parallel corpus will be given multiple translations which will be used at random, resulting in illogical sentences to be generated in the synthetic corpus.

3.2 Synthetic Data Generation

This bilingual lexicon is then used to word-for-word back-translate an in-domain corpus from the target language into the source language. Each word in the monolingual target data is looked up in the lexicon and translated into the corresponding source language word. If a word is not found in the lexicon it remains unchanged. This is done for the entire text resulting in an in-domain monolingual corpus in the source language. These two corpora are then used as a pseudo-parallel corpus of training data. This process is shown in Figure 2.

The main advantage of this synthetic data generation technique is that it can be used to take massive in-domain monolingual datasets that are publicly available and translate them into the source language. DALI aims to improve the in-domain lexical accuracy of the model trained on this data. This is achieved through the word-for-word back-translation used by explicitly identifying the correct translation for in-domain words. The lexical accuracy

Dataset	Sentence Pairs
WMT22 (EN-XH)	5 928 260
PriMock47	3 875
Flores	997
Blocker Evaluation Dataset	590
MeMaT-Clinical	181

Table 1: Number of sentence pairs for each dataset.

is also improved by using in-domain monolingual data. This data will be more lexically diverse than any available in-domain parallel data which will be much smaller than the monolingual data.

On the other hand, this synthetic data is limited by the quality of the bilingual lexicon used, as discussed previously. If there are many mistakes in the lexicon they will be carried through to the synthetic data leading to nonsensical sentences being generated. Another issue is if a word is OOV it will not be changed, meaning that there may be target language words left in the source text after back-translation. This will cause the model to think the words are identical in both languages, which may not be true.

3.3 Uses for This Data

The synthetic data generated can be used as both the source and target language data, allowing for models to be trained to translate in either direction. Since the original data will likely be of a higher quality than the synthetic data, training a model with the synthetic data as the source language will likely result in a better performing MT model. This is because the target language data will be of a higher quality, teaching the model to generate higher quality translations.

You could train a model from scratch using this data however this would not be ideal. This is because the data is not of a high quality grammatically, meaning the model would generate grammatically incorrect sentences. The best way to use this data is to fine-tune a pre-trained model with it. This is because the model will already know how to create grammatically correct sentences, compensating for the lack thereof in the training data, and the DALI [14] generated data should improve the lexical accuracy of the translations.

4 METHODS AND MATERIALS

4.1 Data

Several datasets were used during this project. Table 1 presents the number of sentence pairs in each dataset to show their individual sizes.

4.1.1 Lexicon Induction Datasets. The Workshop on Machine Translation (WMT) is an annual conference held with the focus on researching machine translation. The conference held in 2022 (WMT22) had a shared task for African MT and released datasets for it [1]. In this project the English-isiXhosa subset was used, which contains 8.69 million sentence pairs, as a general domain dataset. A refined subset of these sentences was used in this project.

The Hugging Face repository [9] included three scores associated with each sentence pair. The LASER score measures the semantic similarity of the sentences to ensure they convey the same meaning [3]. The source and target LID scores measure the likelihood that the given sentences are in the expected language [21]. Only the sentence pairs with a LASER score between 0.8 and 1.0 and a source and target LID score of above 0.95 were used, resulting in approximately 5.9 million sentence pairs. This was done to reduce the amount of noise in the data, such as non-isiXhosa sentences.

This dataset was used to create the general-domain bilingual lexicon for use in the DALI process. This was done using fast-align [7]. This is shown in Figure 1 where the WMT22 data is input to fast-align and the lexicon is output. This dataset will be referred to as the WMT22 dataset.

Fast-align takes a parallel corpus of sentences and outputs the most likely mappings of source to target language words. For each word in a source sentence a weight is assigned to every word in the target sentence, representing the strength of their association. The target word with the largest weight for a source word is then aligned to that word in the output. This is an improvement over the IBM model 2, producing results faster and more accurately, as it eliminates the need for complex probability calculations, replacing them with the weight system.

The bilingual medical dictionary used by Blocker et al. (2025) was also used during the lexicon generation [6]. This dictionary was used to adapt the general-domain lexicon to the medical domain. This is shown in Figure 1 where it is referred to as the small in-domain lexicon.

4.1.2 Training Dataset. PriMock57 is a collection of 57 mock doctor-patient consultations recorded in English [18]. This dataset was used as the monolingual corpus of data which was word-for-word back-translated into synthetic isiXhosa data, thereby creating a pseudo-parallel corpus, as seen in Figure 2. This back-translation was done using a script in the DALI Git Hub repository [14].

Blocker et al. (2025) randomly selected ten of these consultations and used them to create their dataset. Since this project will be using that dataset to evaluate the models, they could not be trained on those ten consultations. Therefore, before using PriMock57 those ten consultations were removed resulting in a truncated dataset which will be referred to as PriMock47. This is done to avoid contamination between the training and test sets, to ensure accurate evaluation scores.

4.1.3 Validation Datasets. High-quality validation sets are rare for low-resource languages. Because of this two approaches for validating the hyperparameters were used. Both a general domain and a medical domain validation set were used to fine-tune the hyperparameters to better evaluate the models and how they perform in both domains. The following validation sets were used for hyperparameter tuning:

(1) The Medical Machine Translation (MeMaT) project contains an English-isiXhosa parallel corpus of data and is used in this project for evaluation. There are various subsets of data but only the medical subset was used [16] as an in-domain validation set. This dataset will be referred to as MeMaT medical.
(2) The Flores-200 dataset was created by Meta's NLLB team [21] and builds on the Flores-101 dataset [13] to include

more languages. This project uses this dataset as a general domain evaluation set and it will be referred to as Flores.

4.1.4 Test Dataset. Blocker et al. (2025) presented a new dataset that took ten randomly selected consultations from PriMock57 and adapted them to a South African context [6]. This was done by having South African actors read the consultations in both English and isiXhosa. The isiXhosa consultations were translated from English by a professional translator. The resulting dataset contained 590 parallel sentence pairs. This used as the test set because it is human-translated, making it a reliable evaluation dataset. This dataset will be referred to as the Blocker dataset.

4.2 DALI Set-up

This subsection will describe how the DALI pipeline, discussed in Section 3, was used in this project to generate the data used for training.

(1) The word-alignment tool fast-align [7] was run on the English-isiXhosa subset of the WMT22 dataset [9] to generate the general domain English-isiXhosa lexicon. In Figure 1 the parallel general domain corpus refers to WMT22.
(2) The Blocker medical dictionary was added to this lexicon to adapt it to the medical domain. In Figure 1, the Blocker dictionary is the small in-domain lexicon. This new medical bilingual lexicon covers both basic grammar as well as specialised medical terminology.
(3) The PriMock47 dataset was word-for-word back-translated using this lexicon, producing the synthetic isiXhosa translations of the sentences. In Figure 2 PriMock47 is the in-domain target language corpus.
(4) The original English PriMock47 and the synthetic isiXhosa PriMock47 are then used as a pseudo-parallel corpus of training data.

4.3 Alternative Approaches

As part of the overall study, team members pursued alternative approaches to improving MT for isiXhosa in the medical domain.

The first approach was to train a model without adapting it to the medical domain. This was done by fine-tuning NLLB-200 [10] on the WMT22 dataset [9] to improve its performance on isiXhosa translation further. Then a model was trained from scratch using the WMT22 dataset for general domain English-isiXhosa translation. These two models were used as baselines for comparison with the models adapted to the medical domain.

The second approach was to create synthetic data through back-translation to fine-tune NLLB-200. The same PriMock47 data was used for this purpose. The English PriMock47 data was translated into isiXhosa using NLLB-200 to create a parallel corpus of data that could be used for fine-tuning.

The final results of these models will be discussed in the results section to see how they performed in comparison to the final results of the models trained in this project.

4.4 Evaluation

4.4.1 Reference-Based Metrics. Three of the most common evaluation metrics used for MT models are BLEU [25], chrF [26] and chrF++ [27]. These scores were generated using sacrebleu [28], a python library used for evaluating MT models. Each of these metrics were used to evaluate the models that were trained in this project.

BLEU evaluates the quality of machine translations by comparing them to reference translations done by human translators. This comparison is done by calculating a modified n-gram precision metric between the candidate and reference translations.

chrF differs from BLEU by using a character level analysis of the translation whereas BLEU analyses it on a word level. This makes the chrF metric more sensitive and therefore it is often closer to human evaluations than BLEU.

chrF++ can be seen as a combination of BLEU and chrF as it includes both character and word n-grams which are averaged to get a more precise score. This metric is often closer to human evaluations than both BLEU and chrF. chrF and chrF++ are better metrics for evaluating isiXhosa translations as they often involve sub-words that would not be taken into account by BLEU.

4.4.2 Medical Terminology. The number of medical terms used in the PriMock47 dataset that were also in the Blocker medical dictionary were counted. Then the predictions of the best performing models were evaluated to calculate how many medical terms were translated incorrectly as a percentage. This health term error rate was calculated to see how well these models actually perform in the medical domain.

4.5 Hyperparameter Tuning

A hyperparameter sweep was performed to fine-tune NLLB-200 [10] with the pseudo-parallel corpus to get a range of results. This was done using the transformers python library, developed by Hugging Face [11], which contains pre-trained models including NLLB-200 as well as tools for fine-tuning those models. A total of 382 models were trained, half of those models translate from English → isiXhosa and half translate from isiXhosa → English. The 191 models trained for each direction used the same hyperparameter search space which is shown in Table 2. The results of evaluating these models showed which hyperparameters were the most effective in fine-tuning NLLB-200.

The models were evaluated on the Flores-200 [21] and MeMaT medical [16] validation sets. One general domain (Flores) and one medical domain (MeMaT) dataset was used to evaluate the models' performance in both of these domains. The models that performed the best on each of these validation sets were identified by finding

Hyperparameter	Search Space
Learning Rate	1e-6, 1e-5, 2e-5, 3e-5, 5e-5, 1e-4
Effective Batch Sizes	32, 64
Num. Epochs	3, 5
Warmup Ratios	0.0, 0.1
Weight Decays	0.0, 0.01
Label Smoothings	0.0, 0.1

Table 2: Hyperparameter search space used for sweep

Hyperparameter	En→Xh	Xh→En
Learning Rate	1e-9, 1e-8, 5e-8, 1e-7, 5e-7	
Effective Batch Sizes	64	32
Num. Epochs	3	5
Warmup Ratios	0	0.1

Table 3: Hyperparameter search space used for the additional, smaller sweep. The same learning rates were tested for both directions.

Model ID	Learning Rate	Batch Size	Epochs	Warmup Ratio
En→Xh-Med	1e-6	64	3	0.0
En→Xh-Gen	1e-6	64	3	0.1
Xh→En-Med	1e-7	32	5	0.1
Xh→En-Gen	5e-7	32	5	0.1

Table 4: Hyperparameters of the four best-performing models. These were selected from the search space defined in Table 2. Weight Decays and Label Smoothing have been omitted as they were both 0.0 for all models.

the highest BLEU scores [25]. The hyperparameters of these models were then used as a starting point for an additional search.

This additional search kept the same hyperparameters as the best performing models in each direction except for the learning rates. The learning rates had the largest effect on a model's performance, therefore additional learning rates were explored. It was observed that all the best performing models in the initial sweep used the smallest learning rate (1e-6), so smaller learning rates were explored to increase performance. Table 3 shows the hyperparameters used for both directions.

5 RESULTS AND DISCUSSION

5.1 Hyperparameter Search

After evaluating all of the models on Flores-200 [21] and MeMaT medical [16], the top four performing models were chosen to evaluate on the Blocker dataset. These are two Eng→Xho translation models and two Xho→Eng translation models and the hyperparameters used to train them can be found in Table 4. These models were chosen by finding the models with the highest BLEU score [25] for each evaluation dataset. The models have been named accordingly, with "-Med" denoting the best performing model on the MeMaT medical dataset and "-Gen" denoting the best performing Flores model. The results that are presented in Table 5 illustrate how these models performed for both datasets.

Even though the MeMaT medical dataset is relatively small, which is shown in table 1, the results are still important as they give an indication of how well the models perform in the target domain. The results in Table 5 show that the models perform better on MeMaT for Eng→Xho translation than on Flores with a difference of 3.57 BLEU. For Xho→Eng translation, the opposite is true as both BLEU scores for Flores are better than those for MeMaT.

Figure 3 shows that the smaller the learning rate the higher the average BLEU score achieved by the models until about 1e-6 for

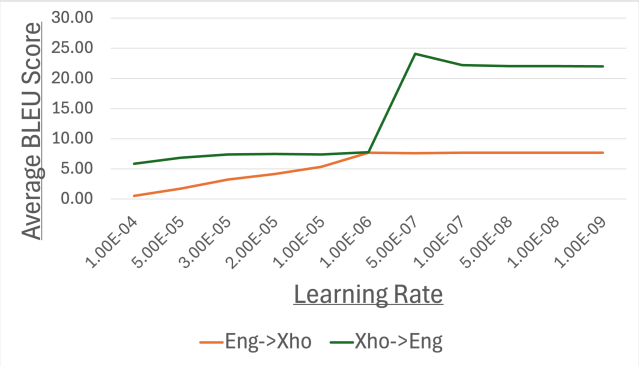

Figure 3: Average BLEU scores of models evaluated on Flores and MeMaT medical with different learning rates.

Eng→Xho translation and 5e-7 for Xho→Eng translation. After these points there is a slight decrease and then a plateau in the scores achieved. This shows the optimal learning rate for each direction when fine-tuning NLLB [10] on the PriMock47 dataset.

5.2 Comparing DALI Against NLLB

The final step for assessing these models is to evaluate them on the Blocker dataset [6] to see how well they perform for isiXhosa translation in the medical domain. The results of these experiments can be seen in Table 6 where they are compared to the results of evaluating the baseline NLLB-200 model [10] on the Blocker dataset.

For Eng→Xho translation the baseline NLLB outperformed the fine-tuned models with a BLEU score [25] of 10.84, attaining approximately one BLEU point above the fine-tuned models. This is likely due to overfitting and catastrophic forgetting by fine-tuning such a large model on a relatively small dataset such as PriMock47. The hyperparameters used could have weakened the models' ability to use general translation rules, instead favouring the specific translations given to it during fine-tuning causing a drop in performance. Another, more simple explanation for this drop in performance is that the synthetic data generated using the DALI process [14] is not good enough for fine-tuning.

For Xho→Eng translation, both fine-tuned models outperformed the baseline model with a top BLEU score of 19.03 (1.11 more than NLLB-200). This improvement demonstrates that the model has successfully been adapted to the medical domain. After fine-tuning, the model is now able to translate specialised terminology more accurately than before.

These results show that DALI has improved the performance of NLLB for Xho→Eng translation but not for Eng→Xho translation. This can be explained by the fact that generating isiXhosa is more difficult than generating English since the original NLLB model is trained on much more English data. It could also be due to the highly agglutinative nature of isiXhosa compared to the morphologically simple nature of English. There is more room for error when constructing large complex isiXhosa words from English, making Eng→Xho translation more difficult. Conversely, de-constructing

	FLORES-200			MEMAT Medical		
Model ID	**BLEU**	**chrF**	**chrF++**	**BLEU**	**chrF**	**chrF++**
English to isiXhosa						
En→Xh-Med	5.98	50.39	42.49	**9.63**	**47.14**	**43.07**
En→Xh-Gen	**6.06**	**50.45**	**42.55**	**9.63**	**47.14**	**43.07**
isiXhosa to English						
Xh→En-Med	22.24	52.14	49.19	**22.19**	**40.46**	**40.03**
Xh→En-Gen	**26.31**	**53.78**	**51.38**	21.82	39.67	39.32

Table 5: Evaluation results of the best-performing models on the FLORES-200 and MEMAT medical test sets. The best score for each metric within a direction is highlighted in bold.

Model ID	**BLEU**	**chrF**	**chrF++**
English to isiXhosa			
En→Xh-Med	9.61	47.57	41.93
En→Xh-Gen	9.72	47.71	42.08
NLLB-200	**10.84**	**48.31**	**42.47**
isiXhosa to English			
Xh→En-Med	18.1	40.58	38.58
Xh→En-Gen	**19.03**	**41.01**	**39.13**
NLLB-200	17.92	40.55	38.53

Table 6: Evaluation results of top models and NLLB-200 on the Blocker dataset.

these complex words into their distinctive morphemes and translating those into English is a much simpler task, meaning Xho→Eng translation is easier.

5.3 Comparing DALI Against Baselines and Back-Translation

In this section the results of the alternative approaches explained in section 4.3 will be discussed in comparison to the results of the models in this project. The results of these experiments are presented in Table 7.

The first alternative approach was to fine-tune NLLB-200 [10] on the WMT22 dataset [9] and to train a new model from scratch using the same dataset. These models serve as a set of baseline models as they are not fine-tuned on any medical domain data.

The second alternative approach was to fine-tune NLLB-200 on data that was synthetically generated using back-translation (referred to as the back-translation models). The results of these models indicate that this approach to generating synthetic data is more effective than DALI [14] as they greatly outperform both the DALI models and the baseline NLLB model. The performance increase between these models and the baseline NLLB is notable with a difference in BLEU [25] of 7.02 for Eng→Xho and 16.54 for Xho→Eng .

The significant performance discrepancy between the back-translation (BT) models and the DALI models (8.14 for Eng→Xho and 15.45 for Xho→Eng) likely emanates from the different data-generation processes. The data generated by DALI utilises word-for-word back-translation which causes the natural grammatical structure of the sentences to be lost. Conversely, the synthetic models use data that was generated by NLLB-200, ensuring that the grammatical structure of the sentences was preserved.

The NLLB model fine-tuned on WMT22 (WMT22-NLLB) performed the worst overall with a BLEU score of 2.08 and 5.84 for Eng→Xho and Xho→Eng translation respectively. This low performance could be because WMT22 is not of a high enough quality to be used for fine-tuning. The dataset could have caused the model to overwrite some it's basic translation rules with incorrect ones found in the dataset. This in turn would lead to the model producing incorrect translations. The model trained from scratch performed moderately better with BLEU scores of 4.6 for Eng→Xho and 11.4 for Xho→Eng . This slight performance increase over WMT22-NLLB could be because that model might have been confused with many different translation rules while the model trained from scratch only has the one set of rules, allowing for more accurate translations.

5.4 Medical Terminology Analysis

Another metric used to evaluate these models is the health term error rate which shows how accurately each model predicted specialised medical terminology in the Blocker dataset. The results of these evaluations are shown in Table 8. For Eng→Xho translation the baseline model had the lowest error rate in all categories (except for condition where it tied with both other models at 66.67%) suggesting again that the small fine-tuning dataset caused the model to over fit certain terms. For Xho→Eng translation the Xh→En-Gen model outperforms the baseline in all categories except for treatment, where its error rate is 3.7% higher than the baseline.

6 CONCLUSIONS

This project investigated efficacy of the DALI process for fine-tuning NLLB-200 for the isiXhosa medical domain. The results of the experiments indicate that DALI-generated data is ineffective for this purpose. For Eng→Xho translation, the fine-tuned models underperformed the NLLB baseline, while for Xho→Eng translation there were only marginal performance gains. These outcomes indicate that the synthetic data lacks the grammatical

Model	**BLEU**	**chrF**	**chrF++**
English to isiXhosa			
BT-Gen (En→Xh)	**17.86**	47.42	**43.8**
WMT22-NLLB	2.08	5.92	3.2
Scratch	12.89	39.90	36.29
DALI Best Scores	9.72	47.71	42.08
NLLB-200	10.84	**48.31**	42.47
isiXhosa to English			
BT-Gen (Xh→En)	**34.46**	**48.06**	**48.05**
WMT22-NLLB	5.84	31.05	26.49
Scratch	14.37	31.43	30.57
DALI Best Scores	19.03	41.01	39.13
NLLB-200	17.92	40.55	38.53

Table 7: Results of the baseline models, synthetic models, best performing DALI models and the baseline NLLB models.

Model ID	Overall	Anatomy	Condition	Treatment
English to isiXhosa				
En→Xh-Med	44.07%	35.62%	**66.67%**	37.04%
En→Xh-Gen	44.07%	35.62%	**66.67%**	37.04%
NLLB-200	**43.22%**	**34.93%**	**66.67%**	**33.33%**
isiXhosa to English				
Xh→En-Med	24.12%	12.56%	55.74%	**48.15%**
Xh→En-Gen	**20.90%**	**8.07%**	**54.10%**	51.85%
NLLB-200	23.79%	12.11%	55.74%	**48.15%**

Table 8: Health Term Error Rate (%) for the best performing models and the NLLB-200 baseline. The best (lowest) score in each category is highlighted in bold.

and semantic knowledge required for effective fine-tuning. This is caused by the word-for-word back-translation process followed in the DALI pipeline. Furthermore, the baseline model's superior performance for health term translation suggests that fine-tuning on this small, low-quality corpus may have caused catastrophic forgetting. This led to the model to lose some of the powerful, pre-trained knowledge that is its primary advantage.

The foremost limitation of this project is the quality of the DALI-generated data. Reducing the noise in the bilingual lexicon would improve the performance of the models fine-tuned on this data. Alternatively, combining PriMock47 with a real parallel corpus could also improve performance. The combination of the synthetic and real data would prevent catastrophic forgetting and preserve the grammatical structure of natural sounding sentences. Finally, using human evaluations instead of the automatic evaluations done in this project would provide a better understanding of how usable the models' translations are in a real medical context.

These findings serve as a critical data point on the limitations of lexicon-based data generation for low-resource languages in specialised domains. This research provides a cautionary result for using DALI in a low-resource setting. By effectively demonstrating the shortfalls of this approach, this project helps steer future research toward more effective techniques for bridging the data gap in low-resource, specialised MT.

REFERENCES

[1] Adelani, D., Alam, M. M. I., Anastasopoulos, A., Bhagia, A., Costa-jussà, M. R., Dodge, J., Faisal, F., Federmann, C., Fedorova, N., Guzmán, F., et al. 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), pp. 773–800.

[2] Aharoni, R., Johnson, M., and Firat, O. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089* (2019).

[3] Artetxe, M., and Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics 7* (09 2019), 597–610.

[4] Bahdanau, D. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[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] Dyer, C., Chahuneau, V., and Smith, N. A. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Atlanta, Georgia, June 2013), L. Vanderwende, H. Daumé III, and K. Kirchhoff, Eds., Association for Computational Linguistics, pp. 644–648.

[8] Edunov, S., Ott, M., Auli, M., and Grangier, D. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381* (2018).

[9] et al., N. No language left behind: Scaling human-centered machine translation. https://huggingface.co/datasets/allenai/wmt22_african, 2022.

[10] Face, H. facebook/nllb-200-distilled-600m. https://huggingface.co/facebook/nllb-200-distilled-600M, 2025.

[11] Face, H. Mms model documentation - transformers. https://huggingface.co/docs/transformers/main/en/model_doc/mms, 2025.

[12] 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).

[13] 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, 2021.

[14] Hu, J., Xia, M., Neubig, G., and Carbonell, J. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 2989–3001.

[15] 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.

[16] Keet, M., Mahlaza, Z., Heafield, K., Birch, A., Pal, P., and Kwebulana, N. Memat: Medical machine translation corpus (isixhosa–english). https://github.com/mkeet/MeMaT, 2018. Corpus developed collaboratively by the University of Cape Town and the University of Edinburgh as part of the EPSRC-funded MeMaT project (WT5518939). Released under a CC-BY license.

[17] Koehn, P., and Knowles, R. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872* (2017).

[18] Korfiatis, A. P., Moramarco, F., Sarac, R., and Savkov, A. Primock57: A dataset of primary care mock consultations. *arXiv preprint arXiv:2204.00333* (2022).

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

[20] 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.

[21] NLLB Team, Marta R. Costa-jussà, J. C. O. M. E. K. H. K. H. E. K. J. L. D. L. J. M. A. S. S. W. G. W. A. Y. B. A. L. B. G. M. G. P. H. J. H. S. J. K. R. S. D. R. S. S. C. T. P. A. N. F. A. S. B. S. E. A. F. C. G. V. G. F. G. P. K. A. M. C. R. S. S. H. S. J. W. No language left behind: Scaling human-centered machine translation, 2022.

[22] Nyoni, E., and Bassett, B. A. Low-resource neural machine translation for southern african languages. *arXiv preprint arXiv:2104.05579* (2021).

[23] Och, F. J., and Ney, H. A systematic comparison of various statistical alignment models. *Computational linguistics 29*, 1 (2003), 19–51.

[24] Omniscien Technologies. What is neural machine translation?, n.d.

[25] 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.

[26] Popović, M. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation* (2015), pp. 392–395.

[27] Popović, M. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation* (2017), pp. 612–618.

[28] Post, M. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (Belgium, Brussels, Oct. 2018), Association for Computational Linguistics, pp. 186–191.

[29] Statistics South Africa. South africa's evolving cultural landscape: A 26-year transformation, Mar. 2025.

[30] 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).

[31] 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.